

Studies in Systems, Decision and Control 483

Nguyen Ngoc Thach  
Vladik Kreinovich  
Doan Thanh Ha  
Nguyen Duc Trung *Editors*

# Optimal Transport Statistics for Economics and Related Topics

 Springer

Series Editor

Janusz Kacprzyk, *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*



The series “Studies in Systems, Decision and Control” (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control—quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Nguyen Ngoc Thach · Vladik Kreinovich ·  
Doan Thanh Ha · Nguyen Duc Trung  
Editors

# Optimal Transport Statistics for Economics and Related Topics

*Editors*

Nguyen Ngoc Thach  
Ho Chi Minh University of Banking  
Ho Chi Minh City, Vietnam

Vladik Kreinovich  
Department of Computer Science  
Texas A&M University  
El Paso, TX, USA

Doan Thanh Ha  
Ho Chi Minh City University of Banking  
Ho Chi Minh City, Vietnam

Nguyen Duc Trung  
Ho Chi Minh City University of Banking  
Ho Chi Minh City, Vietnam

ISSN 2198-4182

ISSN 2198-4190 (electronic)

Studies in Systems, Decision and Control

ISBN 978-3-031-35762-6

ISBN 978-3-031-35763-3 (eBook)

<https://doi.org/10.1007/978-3-031-35763-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license  
to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Why optimal transport statistics? Of course, transportation is an important part of economic activity, and solutions to transportation problems are important part of econometrics. However, the main reason why optimal transport techniques are useful in econometric goes beyond applications to transportation. The main reason is that the analysis of optimal transport problems led to the development of new economically reasonable ways to gauge the difference between probability distributions. Taking into account that in economics, we can only make predictions about probabilities of different events, the resulting measures of difference help analyze how close are the predicted probabilities to the actual ones.

This volume emphasizes techniques of optimal transport statistics, but it also describes and uses other econometric techniques, ranging from more traditional statistical techniques to more innovative ones such as quantiles (in particular, multidimensional quantiles), maximum entropy approach, and machine learning. Applications range from general analysis of GDP growth, stock market, and consumer prices to analysis of specific sectors of economics (construction, credit and banking, energy, health, labor, textile, tourism, international trade) to specific issues affecting economy such as bankruptcy, effect of COVID-19 pandemic, effect of pollution, effect of gender, cryptocurrencies, and the existence of shadow economy. Papers presented in this volume also cover data processing techniques, with economic and financial application being the unifying theme.

This volume shows what has been achieved, but even more important are remaining open problems. We hope that this volume will:

- inspire practitioners to learn how to apply state-of-the-art techniques, especially techniques of optimal transport statistics, to economic and financial problems, and
- inspire researchers to further improve the existing techniques and to come up with new techniques for studying economic and financial phenomena.

We want to thank all the authors for their contributions and all anonymous referees for their thorough analysis and helpful comments.

The publication of this volume is partly supported by the Ho Chi Minh University of Banking, Vietnam. Our thanks to the leadership and staff of this university, for providing crucial support. Our special thanks to Prof. Hung T. Nguyen for his valuable advice and constant support.

We would also like to thank Prof. Janusz Kacprzyk (Series Editor) and Dr. Thomas Ditzinger (Senior Editor, Engineering/Applied Sciences) for their support and cooperation with this publication.

February 2023

Nguyen Ngoc Thach  
Vladik Kreinovich  
Doan Thanh Ha  
Nguyen Duc Trung

# Contents

An Invitation to Multivariate Quantiles Arising from Optimal Transport Theory .....	1
<i>Hung T. Nguyen</i>	
The Unfulfilled Quest for Discovering Cause from Probability .....	38
<i>William M. Briggs</i>	
Optimal Transport for Counterfactual Estimation: A Method for Causal Inference .....	45
<i>Arthur Charpentier, Emmanuel Flachaire, and Ewen Gallic</i>	
Three Applications of Measure Transportation in Statistical Inference .....	90
<i>Marc Hallin</i>	
Extending the A Priori Procedure for Estimating Location Parameter Under Multivariate Skew Normal Settings .....	107
<i>Ziwei Ma, Tonghui Wang, S. T. Boris Choy, Zheng Wei, and Xiaonan Zhu</i>	
A Note on Cournot-Nash Equilibria and Optimal Transport Between Unequal Dimensions .....	117
<i>Luca Nenna and Brendan Pass</i>	
Stacking Regression for Time-Series, with an Application to Forecasting Quarterly US GDP Growth .....	131
<i>Erkal Ersoy, Haoyang Li, Mark E. Schaffer, and Tibor Szendrei</i>	
Maximum Entropy Learning with Neural Networks .....	150
<i>Woraphon Yamaka</i>	
Robustness of Multi-criteria Nash Equilibrium Based on Vectorial Rationality Function .....	163
<i>Urairat Deepan, Parin Chaipunya, and Poom Kumam</i>	
Why Quantiles Are a Good Description of Volatility in Economics: An Alternative Explanation .....	169
<i>Laxman Bokati, Olga Kosheleva, Vladik Kreinovich, and Kittawit Autchariyapanitkul</i>	
Hawthorne Effect: An Explanation Based on Decision Theory .....	174
<i>Sofia Holguin, Vladik Kreinovich, and Phuong Hoang Nguyen</i>	

Fair Bankruptcy Solutions Under Interval Uncertainty .....	178
<i>Uyen Pham, Olga Kosheleva, and Vladik Kreinovich</i>	
Economy-Related Emotional Attitudes Towards Other People: How Can We Explain Them? .....	186
<i>Christopher Reyes, Vladik Kreinovich, and Chon Van Le</i>	
COVID-19 and Short-Run Survival in the Service Sector: Evidence from the Tourism Economy .....	193
<i>Surapot Baiya, Pithoon Thanabordeekij, and Paravee Maneejuk</i>	
How Non – Interest Income Matters for Operation Efficiency? A Bayesian Analysis of Vietnam Banks .....	211
<i>Bui Dan Thanh, Doan Thanh Ha, and Pham Thi Hong Nhung</i>	
Cryptocurrency Portfolio Management Based on Usage Characteristics Criteria Applying R-Vine Copula .....	235
<i>Terdthiti Chitkasame, Pichayakone Rakpho, and Nachattapong Kaewsompong</i>	
Does Debt Affect Profitability of Construction Companies in Vietnam? A Bayesian Approach .....	248
<i>Bui Dan Thanh and Nguyen Ngoc Huyen</i>	
Determinants of Small and Medium Enterprises' Capital Intensity: The Case in Vietnam .....	264
<i>Nhan Truong Thanh Dang, Van Dung Ha, and Van Tung Nguyen</i>	
Labor Productivity: Does Export Matter for Vietnamese Small and Medium Enterprises? .....	274
<i>Dang Nhan Truong Thanh, Ha Van Dung, and Nguyen Van Tung</i>	
Income and Consumption Patterns of Sri Lankan Senior Citizens and Subsequent Impact on Policies and Transportation .....	286
<i>Shanika Madushani Jayathunga and Gnanadarsha Sanjaya Dissanayake</i>	
Bayesian Consideration for Analyzing Employee's Motivation: Evidence from Vietnam .....	299
<i>Bui Huy Khoi and Nguyen Ngoc Thach</i>	
The Roles of Grassroots Government and Associations Versus Internet Access in Households' Income in Vietnam .....	315
<i>Chon Van Le and Thuong Thi Vu</i>	

Predicting the Impact of Covid Pandemic on the Relationship Between Logistics Activities and Business Performance: A PLS-SEM Approach .....	327
<i>Le Thi Phuong Thanh, Le Thi Phuong Thao, and Tong Viet Bao Hoang</i>	
Consumption Expenditure Comparison Among Vulnerable Households in Thailand .....	345
<i>Supanika Leurcharusmee and Anaspree Chaiwan</i>	
The Presence of Child and Spouse in the Household and Labor Market Opportunities of Male and Female Workers in Thailand .....	358
<i>Supanika Leurcharusmee and Anaspree Chaiwan</i>	
Link Between Renewable and Non-renewable Energy Consumption and Co2 Emissions: A Monte-Carlo Simulation Study .....	376
<i>Phan Thi Lieu and Nguyen Ngoc Thach</i>	
Machine Learning Applications on Box-Office Revenue Forecasting: The Taiwanese Film Market Case Study .....	384
<i>Shih-Hao Lu, Hung-Jen Wang, and Anh Tu Nguyen</i>	
The War on the Shadow Economy in Southeast Asia: A New Contribution from Inclusion of LGBT People .....	403
<i>Duong Tien Ha My, Nguyen Ngoc Thach, Phan Thi Minh Hue, and Nguyen Van Diep</i>	
Bayesian Hierarchical Mix-Effects Approach to Impacts of Air Pollution and Economic Growth on Private Health Care Expenditure .....	417
<i>Bui Hoang Ngoc and Nguyen Ngoc Thach</i>	
Market Share Forecast of Vietnam and of the World's Leading Textile and Garment Exporters by VAR Bayesian Model .....	427
<i>Nguyen Thi Ngoc Diep, Tran Quang Canh, and Nguyen Ngoc Thach</i>	
The Impact of Global Value Chain Integration on Export: Evidence from Vietnam .....	440
<i>Nguyen Thi Ngoc Diep, Tran Quang Canh, and Nguyen Ngoc Thach</i>	
A Hybrid Model Based on ARIMA and Artificial Neural Network to Forecast Consumer Price Index: The Case of Vietnam .....	449
<i>Thi Thanh Huyen Le and Tien Nhat Nguyen</i>	
A Bayesian Approach to Determinants of Capital Structure of Listed Construction Firms in Vietnam .....	465
<i>Nguyen Duc Trung, Nguyen Ngoc Thach, and Bui Dan Thanh</i>	



Determinant of Capital Adequacy Ratio: Evidence from Commercial Banks in Vietnam .....	480
<i>Nguyen Thi Nhu Quynh and Nguyen Duc Trung</i>	
Impact of Managers' Gender Difference on Firms' Liability in Vietnam .....	498
<i>Van Tung Nguyen, Nhan Truong Thanh Dang, Van Dung Ha, and Thi Anh Tuyet Le</i>	
Contagion Effects Among Commodity Markets and Securities Markets During the Conflict Between Russia and Ukraine: The Dynamic Conditional Correlation Approach .....	513
<i>Sunisa Phaimekha and Worrawat Saijai</i>	
Net Interest Margins of Vietnamese Commercial Banks: What Really Affects? .....	523
<i>Nam Hai Pham and Thuy Kieu Thi Vo</i>	
Credit Growth: An Investigation of Vietnamese Commercial Banks .....	533
<i>Nam Hai Pham, Nguyen Ngoc Thach, Tuan Van Ngo, and Tri Minh Hoang</i>	
Forecasting the Exchange Rate for the Thai Baht Against the Chinese Yuan by Using a Genetic Algorithm-Based Subset Autoregressive Integrated Moving Average Model .....	544
<i>Tassathorn Poonsin, Vayu Thanomsing, Thanakorn Thunjang, and Worrawate Leela-apiradee</i>	
Impacts of Countermeasure Program on the Covid-19 Pandemic in Asian Countries .....	560
<i>Worrawat Saijai and Sukrit Thongkairat</i>	
Correlation Between Foreign Ownership and Liquidity Risk .....	574
<i>Nguyen Ngoc Thach, Bui Dan Thanh, and Le Thi Lan</i>	
Impacts of Financial Development on Vietnamese Commercial Banks' Lending Mechanisms of Monetary Policy Pass-Through: Bayesian Analysis .....	588
<i>Thi Thu Hong Dinh, Thanh Phuc Nguyen, and Ngoc Tho Tran</i>	
A Bayesian Binary Logistic Regression Approach to Identifying Factors Affecting the Households' Use Level of Financial Products/Services in Vietnam .....	612
<i>Huong Thi Thanh Tran</i>	

Impacts of Global Pandemics, Financial Crises, and Oil Price Shocks on Japanese Stock Market ..... 627  
*Rongchai Tansuchat and Chaiwat Klinlampu*

Tourism Business Adaption to Survive the Coronavirus Disease-2019 Pandemic in Thailand ..... 637  
*Supareuk Tarapituxwong, Piangtawan Polard, and Namchok Chimprang*

Impacts of Capital Structure on Microfinance Institutions’ Risk: Evidence from Low- and Middle-Income Countries ..... 654  
*Thuy T. Dang, Nguyen Tran Xuan Linh, Hau Trung Nguyen, and Dinh Cong Hoang*

Factors Affecting the Financial Leverage of Vietnam Businesses ..... 667  
*Thi Anh Tuyet Le, Nhan Truong Thanh Dang, Van Dan Nguyen, and Van Tung Nguyen*

Understanding the Nexus Between Emerging Stock Market Volatility and Gold Price Shocks ..... 676  
*Woraphon Yamaka*

How Does Energy Consumption Matter for Economic Growth? A Bayesian Data Analysis ..... 691  
*Nguyen Ngoc Thach and Phan Thi Lieu*

**Author Index** ..... 699



# An Invitation to Multivariate Quantiles Arising from Optimal Transport Theory

Hung T. Nguyen<sup>1,2</sup>(✉)

<sup>1</sup> New Mexico State University, Las Cruces, USA  
hunguyen@nmsu.edu

<sup>2</sup> Chiang Mai University, Chiang Mai, Thailand

**Abstract.** While a variety of new concepts and methods arised from Optimal Transport theory recently in the literature, they are somewhat theoretical for empirical researchers, including statisticians and econometricians. This tutorial paper aims at elaborating on one of these new concepts and methods, namely multivariate quantile functions, in order to invite empirical researchers to take a closer look at this new concept to apply to their empirical works, such as multivariate quantile regression.

**Keywords:** Gradient of convex functions · Multivariate quantiles · Optimal transport · Quantile regression

## 1 Introduction

Motivated by economics issues, in 1942, Kantorovich reformulated (and solved) the unsolved “Optimal Transport” (OT) problem of Gaspard Monge (1781) and got the Nobel Prize in Economics (shared with Koopmans) in 1975, for their contributions to optimal allocation of resources.

Recently, it was “discovered” that OT provides a variety of modern methods for economics. This tutorial paper focuses only on one of these modern methods, namely multivariate quantile functions for quantile regression and related topics.

Mean linear regression models are possible (as it is obvious how to generalize the mean of a random variable to the mean of a random vector) and are useful when dealing with multivariate distribution functions. Now, over 40 years since univariate quantile regression was invented (Koenker and Bassett [5]), can we extend it to multivariate quantile regression in some acceptable way? Of course, like multivariate *mean* linear regressions, multivariate quantile regressions should be very useful in a variety of contexts.

Since there is no total order relation on  $\mathbb{R}^d$  when  $d > 1$ , a direct extension of univariate quantile functions to higher dimensions is hopeless. Thus, in order to obtain a “correct” way to generalize univariate quantile functions, we must look for some other ways. Generalizations of mathematical concepts appear often in mathematics. When Lotfi Zadeh generalized crisp sets to fuzzy sets, he cannot do it directly, so he took

an equivalent definition of a crisp set, namely its indicator function which is a function taking only values 0 and 1, and extend it to the unit interval  $[0, 1]$ . To view Kolmogorov probability theory as a special case of quantum probability theory, we can take an equivalent representation of finite standard probability, namely, identifying a random variable, an event, and a probability measure, as diagonal matrices, and then extend them to arbitrary self adjoint matrices.

Now, in the above spirit, to generalize a univariate quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$ , defined as  $F^{[-1]}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$ , of a real-valued random variable  $X$  with distribution function  $F$ , we look at some appropriate equivalent representation.

Note that, since we are going to derive a characterization of a univariate quantile function in the setting of Optimal Transport theory, we denote it as  $F^{[-1]}$  instead of  $F^{-1}$  to avoid a possible confusion with the set-valued set-function  $T^{-1}(\cdot) : 2^{\mathbb{R}} \rightarrow 2^{[0,1]}$  of a map  $T(\cdot) : [0, 1] \rightarrow \mathbb{R}$ , pushing the uniform probability measure  $du$  on  $[0, 1]$  to a probability measure on  $\mathcal{B}(\mathbb{R})$ , since actually, for  $T = F^{[-1]}$ , we have  $dF = T\#du = du \circ T^{-1}$ !

The first characteristic property of  $F^{[-1]}(\cdot)$  is this: If  $U$  is a random variable, uniformly distributed on the unit interval  $[0, 1]$ , then the random variable  $F^{[-1]}(U)$  has the same distribution  $F$  as  $X$ , which is the basis of simulations. But saying that  $X \stackrel{D}{=} F^{[-1]}(U)$  simply means that the probability “law” of  $X$ , written as  $dF(-\infty, x] = F(x)$ , is the probability measure on  $\mathcal{B}(\mathbb{R})$  obtained from the uniform probability  $du$  on  $\mathcal{B}([0, 1])$  via  $du \circ (F^{[-1]})^{-1}$ , written as  $F^{[-1]}\#du = dF$  (a notation we will use in the context of Optimal Transport Theory), meaning “The transport map  $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$  pushes the probability  $du$  on  $[0, 1]$  forward to the probability  $dF$  on  $\mathbb{R}$ ”.

There is another property of  $F^{[-1]}(\cdot)$ , kind of “hidden”, since we did not use it often.

From the explicit definition of  $F^{[-1]}(u)$ , it is clear that the function  $F^{[-1]}(\cdot)$  is monotone non decreasing on  $\mathbb{R}$ , i.e., if  $x \leq y$  then  $F^{[-1]}(x) \leq F^{[-1]}(y)$ , with is equivalent to: for any  $x, y \in \mathbb{R}$ , we have

$$(F^{[-1]}(x) - F^{[-1]}(y))(x - y) \geq 0$$

and which can be generalized to higher dimensions (needed for our subsequent analysis), as follows. A function  $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is monotone (non decreasing) if, for any  $x, y \in \mathbb{R}^d$ , we have

$$\langle g(x) - g(y), x - y \rangle \geq 0$$

where  $\langle \cdot, \cdot \rangle$  denotes the scalar product on  $\mathbb{R}^d$ .

These two properties characterize the univariate quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$ . Thus, we should expect that a function  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$  is “called” the (multivariate) quantile function of the multivariate distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$  if it possess these two “extended” properties, namely

- (i)  $Q_F$  is monotone non decreasing on  $\mathbb{R}^d$  (in the above equivalent sense),
- (ii)  $Q_F$  pushes the uniform probability  $du$  on  $[0, 1]^d$  forward to the probability  $dF$  on  $\mathbb{R}^d$ , in symbol  $Q_F\#du = dF$ .

Having these requirements, let's see if we can get a candidate for  $Q_F$  in some "simple" way. To simplify the notations, consider the case where the dimension  $d = 2$ .

Thus, let  $F(\cdot) : \mathbb{R}^2 \rightarrow [0, 1]$  a bivariate distribution of the random vector  $X = (X_1, X_2)$ , so that

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = c(F_1(x_1), F_2(x_2))$$

where  $c(\cdot, \cdot) : [0, 1]^2 \rightarrow [0, 1]$  is a bivariate copula capturing the dependence structure between the components of  $X$ .

The mean of the random vector  $X = (X_1, X_2)'$  is defined componentwise, as a mean vector, namely  $EX = (EX_1, EX_2)'$  (transpose).

Can we define bivariate quantile function componentwise?

Let  $Q_F(\cdot) : [0, 1]^2 \rightarrow \mathbb{R}^2$  be defined as, for  $u = (u_1, u_2) \in [0, 1]^2$ ,  $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$ .

a) Monotonicity is satisfied: let  $v = (v_1, v_2)$ , we have

$$\langle Q_F(u) - Q_F(v), u - v \rangle =$$

$$[(F_1^{[-1]}(u_1) - F_1^{[-1]}(v_1))(u_1 - v_1)][(F_2^{[-1]}(u_2) - F_2^{[-1]}(v_2))(u_2 - v_2)] \geq 0$$

since both  $F_1^{[-1]}$ ,  $F_2^{[-1]}$  are monotone.

b) How about  $Q_F \# du \stackrel{?}{=} dF$ ? We have

$$Q_F \# du((-\infty, a] \times (-\infty, b]) = du\{u : Q_F^{-1}((-\infty, a] \times (-\infty, b])\} =$$

$$du\{u : F_1^{[-1]}(u_1) \leq a, F_2^{[-1]}(u_2) \leq b\} = du\{u : u_1 \leq F_1(a), u_2 \leq F_2(b)\} =$$

$$F_1(a)F_2(b) \neq F(a, b)$$

unless  $X = (X_1, X_2)$  has *independent components*, i.e.,  $X_1, X_2$  are independent. This is, in fact, expected since the componentwise definition  $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$  ignores the dependence structure of  $X_1$  and  $X_2$  (given by copulas).

Thus,  $Q_F \# du \neq dF$ , i.e.,  $X \stackrel{D}{\neq} dF$ , in general, meaning that  $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$  is not a good candidate for what we could call a bivariate quantile function, a counterpart of univariate quantile function.

It turns out that a correct candidate for a multivariate quantile function came from an area of mathematics called *Optimal Transport (OT)* theory, in 2016. See Carlier et al. [2, 3].

Let  $\mu, \nu$  be two Borel probability measures on  $\mathbb{R}^d$ ,  $d \geq 1$ . A transport map sending  $\mu$  to  $\nu$  is a map  $T(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , such that  $\nu(\cdot) = \mu \circ T^{-1}(\cdot)$ , i.e.,  $T\#\mu = \nu$ .

Of course, for  $d = 1$ , and  $\mu = du$ , uniform on  $[0, 1]$  and  $\nu = dF$ , for arbitrary distribution function  $F$  on  $\mathbb{R}$ , the quantile function  $F^{[-1]}(\cdot)$  is a transport map sending  $du$  to  $dF$ .

Moreover,  $F^{[-1]}$  is the *unique* monotone transport map (there are other transport maps, but  $F^{[-1]}$  is the only one which is monotone).

Since monotonicity and measure-preserving  $\#$  are concepts which are valid in any dimensions, the question of interest to us is: “Is there a unique monotone transport map on  $\mathbb{R}^d$  for  $\mu = du$ , uniform on  $[0, 1]^d$ , and arbitrary  $\nu = dF$ ?”. If the answer to it is yes, then we get our desired multivariate quantile function! The answer is in fact yes.

*McCann Theorem* (McCann [7]). Let  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$  be an arbitrary multivariate distribution function, then there exists a unique gradient  $\nabla\varphi(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$  of some convex function  $\varphi(\cdot) : [0, 1]^d \rightarrow \mathbb{R}$  ( $\varphi$  is not unique, but  $\nabla\varphi$  is unique) such that  $\nabla\varphi\#du = dF$ , where  $du$  is the uniform probability measure on  $[0, 1]^d$ .

Let’s elaborate a bit on McCann’s Theorem. In dimension 1, let  $\nu = dF$  where  $F$  is the uniform distribution on the interval  $[1, 2]$ , i.e.,

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ x & \text{for } 1 \leq x \leq 2 \\ 1 & \text{for } x > 2 \end{cases}$$

Then we know that  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is  $F^{[-1]}(u) = 1 + u$  which is monotone (non decreasing) since its derivative is positive. It is the derivative of the convex function  $\varphi(u) = \frac{1}{2}(1 + u)^2$ . And of course,  $F^{[-1]}\#du = dF$ .

In dimension 1, the graph of a convex function lies above the tangent at each point  $x$  where the function is differentiable (a convex function is differentiable almost everywhere, with respect to the Lebesgue measure on  $\mathbb{R}$ ), and as such its (a.e.) derivative is monotone non decreasing.

In dimension  $d > 1$ , the gradient is the vector of partial derivatives of the multivariate function. The whole graph of a convex function on  $\mathbb{R}^d$  lies above each tangent hyperplane at each point where it is differentiable, as a consequence, the gradient  $\nabla\varphi$  of the convex function  $\varphi$  is monotone non decreasing in the sense that, for any  $x, y \in \mathbb{R}^d$ , we have

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y \rangle \geq 0$$

McCann’s theorem is an existence theorem, it does not tell us how to obtain explicitly the multivariate quantile function in dimension  $d > 1$ . In other words, it is not a “constructive” theorem. There is much more work to do to get a “constructive” result, and we need it for applications.

It is precisely here that OT comes in.

In 1781, Gaspard Monge [8] considered the following problem. Let  $\mu, \nu$  be two probability measures on  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^3$ , respectively. Let  $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be a “cost” function (of moving the mass  $\mu$  on  $\mathcal{X}$  to the mass  $\nu$  on  $\mathcal{Y}$ , think about “supply and demand”). A transport map  $T(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$  is a map such that  $T\#\mu = \nu$ . The Monge’s problem (MP) is to find a transport map  $T^*$  which is optimal, with respect to the cost  $c$ , in the sense that it minimizes the total cost, i.e.,

$$T^* = \arg \min \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) : T : T\#\mu = \nu \right\}$$

This optimization problem is very difficult to study since the objective function is not linear in  $T$ , and the constraint set is not convex. This is why the problem was dormant for 200 years. Then, in 1942, Kantorovich, motivated by economic problems, solved it, earning him a Nobel Prize in Economics.

The (MP) might not even have solutions! So first of all, when a mathematician faces a such problem, she will enlarge the domain to have solutions, just like considering complex plane for solutions of equations, or extending pure (deterministic) strategies in games to mixed (random) strategies to have Nash equilibrium.

Kantorovich observed that if  $T$  is a solution of (MP), then  $\gamma_T = \mu \circ (I, T)^{-1}$  is a joint probability measure on  $\mathcal{X} \times \mathcal{Y}$  admitting  $\mu$  and  $\nu$  as its marginal measures, i.e.,  $\gamma_T(A \times \mathcal{Y}) = \mu(A)$ , and  $\gamma_T(\mathcal{Y} \times B) = \nu(B)$ , for any  $A \in \mathcal{B}(\mathcal{X})$ ,  $B \in \mathcal{B}(\mathcal{Y})$ . Therefore, the set of all joint probability measures with  $\mu$  and  $\nu$  as marginal measures, denoted as  $\Pi(\mu, \nu)$ , is larger than the set of transport maps in (MP). Note that, by  $(I, T)$ , where  $I$  is the identity map on  $\mathcal{X}$ ,  $I(x) = x$ , we mean the map  $(I, T)(\cdot) : \mathcal{X} \rightarrow \mathcal{X} \times \mathcal{Y}$ ,  $(I, T)(x) = (x, T(x))$ , so that  $(I, T)^{-1}(\cdot) : 2^{\mathcal{X} \times \mathcal{Y}} \rightarrow 2^{\mathcal{X}}$ .

The Kantorovich problem (KP) is this. Find the optimal transport *plan*  $\pi^* \in \Pi(\mu, \nu)$  such that

$$\pi^* = \arg \min \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}$$

If a solution  $\pi^*$  of (KP) is of the form  $\gamma_T = \mu \circ (I, T)^{-1}$ , then, in it,  $T$  is a solution for (MP).

The breakthrough of Kantorovich is this. First of all, unlike (MP), the (KP) always have solutions since  $\Pi(\mu, \nu) \neq \emptyset$  (the product measure  $\mu \otimes \nu$  is in it).

Next, the (KP) seems “solvable” since it avoids the difficulties of (MP): The objective function is linear in  $\pi$ , and the constraint set  $\Pi(\mu, \nu)$  is convex.

As such, the problem can be solved by duality, i.e., changing an “inf” problem to a “sup” problem in which constraints are written as (infinite) inequalities, suitable for using linear programming (invented by Kantorovich himself, 1942, of course, with the help from George B. Danzig). See Villani [11, 12].

The following Sections will explain this program, at least as a gentle introduction, to obtain a constructive theory of multivariate quantiles functions.

## 2 A Closer Look at Quantiles

We are familiar with the notion of (univariate) quantiles when considering order statistics, say, in extreme value theory.

While in practice, we are mainly concerned with distributions of order statistics which are derived solely from the distribution of the population, you may not notice the extremely important role played by the quantile function of the population, although it is derived from the population distribution.

Since the notion of quantile function is essential in various contexts, such as risk analysis, regression models, but until recently is only available for univariate case, i.e., for real-valued random variables, it is desirable to extend it to the multivariate case, i.e., for random vectors, for applications.

The search for such an extension finally arrived (in 2016) by looking closely at the univariate quantile function, triggered by a paper of Brenier [1]. The buzz words in his paper are “polar factorization” and “Monotone rearrangement” of *vector-valued functions*. In fact, this paper first triggered a return to Optimal Transport Theory (OT) since it is precisely in the setting of OT. Specifically, Brenier’s paper is about extension of the above buzz words from *random variables to random vectors*.

Let  $X_1, X_2, \dots, X_n$  be a random sample drawn from a population  $X$ , i.e., the  $X_j$ 's are random variables independent and identically distributed (i.i.d.) as  $X$ . These values of  $X$  are in  $\mathbb{R}$  in any possible order. Suppose we are interested in ordering these observed values of  $X$ , i.e., forming the order statistic  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ , we can just do it!

Can we do it in some more “sophisticated” way? i.e., providing a map that realizes such an ordering.

Note that the order statistic  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  is a *monotone rearrangement* of the values  $X_1, X_2, \dots, X_n$ , i.e., arranging the unordered set  $\{X_1, X_2, \dots, X_n\}$  into the ordered set  $\{X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}\}$ . Of course, this is possible since  $\mathbb{R}$  is totally ordered. Clearly, there is only one such monotone rearrangement.

It is right here that quantile function is related to order statistics. The empirical distribution of the sample is

$$F_n(x) = \frac{1}{n} \sum_{j=1}^n 1_{(-\infty, x]}(X_j)$$

Let the quantile function of  $F_n$  be  $F_n^{[-1]}$ . Then

$$F_n^{[-1]}(u) = X_{(j)} \quad \text{for } u \in \left[ \frac{j-1}{n}, \frac{j}{n} \right)$$

Thus, the quantile function  $F_n^{[-1]}(\cdot)$  (of  $F_n$ ) realizes the monotone rearrangement of the observed values of  $X$ , noting that  $F_n^{[-1]}(\cdot)$  is a monotone non decreasing function.

In fact, we do get a stronger representation than  $F^{[-1]} \# du = dF$ , a weak representation of  $X$ , sufficient for simulation purpose, namely: there exists a random variable  $V$  distributed uniformly on  $[0, 1]$  such that  $X \stackrel{a.s.}{=} F^{[-1]}(V)$ , a *polar factorization* of  $X$ .

Thus, in summary, the quantile function  $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$  provides a polar factorization and a monotone rearrangement for the random variable  $X$ .

The next question is: What is the counterpart of  $F^{[-1]}$  in higher dimensions?, i.e., for  $X$  being a random vector, taking values in  $\mathbb{R}^d$ , with  $d > 1$ .

The answer was given in Brenier’s paper, and later generalized by McCann [7].

Now in the *Text Approximation Theorems of Mathematical Statistics* (Robert J. Serfling, 1980), Serfling started (p. 2–3) as: Let  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$  be the (multivariate) distribution function of a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_d)$ , defined on  $(\Omega, \mathcal{A}, P)$ . The mean of  $\mathbf{X}$  is defined as the mean vector  $E\mathbf{X} = (EX_1, EX_2, \dots, EX_d)$ .



How about quantiles? Well, without explaining why, he considered only the *univariate case* ( $d = 1$ ).

We may ask: Why Serfling did not “consider”, in a parallel way with distribution functions, the notion of quantiles for multivariate distribution functions (but only talked about means of random vectors)? It turns out that this definition of quantile functions for univariate distribution functions is “good” for *simulations*.

The simulation of a univariate random variable  $X$  with distribution function  $F$  is based on the fact that  $X$  and  $F^{[-1]}(U)$ , where  $U$  is the random variable uniformly distributed on the unit interval  $[0, 1]$ , have the same distribution  $F$ . As such, if  $U = u$ , then  $X = x = F^{[-1]}(u)$  is a simulated observation of  $X$ .

**Remark.** As far as simulation of random variables is concerned, we only need the “weak” representation of  $X$ , namely  $X \stackrel{D}{=} F^{[-1]}(U)$ , for any  $F$ , and uniformly distributed  $U$  on  $[0, 1]$ . This representation is termed “weak” since the two random variables  $X$  and  $F^{[-1]}(U)$  are “equal” only in distribution, i.e., having the same distribution, and not necessarily equal almost surely (with probability one) which is a stronger condition. We will see later that *there exists* some random variable  $V \sim U$  such that  $X \stackrel{a.s.}{=} F^{[-1]}(V)$ .

How about *simulations of random vectors*? i.e., how to simulate random vectors when we *do not have* the counterpart notion of quantiles for multivariate distribution functions? In the above mentioned Text, simulation of random vectors is carried out as follows (based on univariate quantile functions only). We elaborate on it in the simple case of dimension two.

Let  $\mathbf{X} = (X_1, X_2)$  with marginal distribution functions  $F_1, F_2$ , and joint distribution  $F$ . The simulation of  $\mathbf{X} = (X_1, X_2)$  is based on the Rosenblatt transformation (1952). Let  $F(x_1, x_2) = F_1(x_1)F(x_2|x_1)$ , define  $(x_1, x_2) \in \mathbb{R}^2 \rightarrow (u_1, u_2) \in [0, 1]^2$  by

$$u_1 = F_1(x_1), u_2 = F(x_2|x_1)$$

The intent is to generate  $u_1, u_2$  independently from a uniform distribution  $du$  on  $[0, 1]$ , then solve the above system of equations (with known marginal and conditional distribution functions of course) to get  $x_1 = F_1^{[-1]}(u_1)$ ,  $x_2 = F_{X_2|X_1}^{[-1]}(u_2)$ , and “view” them as simulated values for  $X_1, X_2$ . This can be justified if, e.g.,  $F_1^{[-1]} \circ F_1 = I$  (identity), and  $X_1, X_2$  obtained this way is distributed as  $F$ . This requires that  $F$  is continuous, so that the Rosenblatt transformation produces  $(U_1, U_2)$  uniformly on  $[0, 1]^2$ .

If  $X_1, X_2$  are independent, i.e.,  $F(x_1, x_2) = F_1(x_1)F_2(x_2)$ , then the simulation process is justified since then the vector  $(F_1^{[-1]}(u_1), F_{X_2}^{[-1]}(u_2))' : [0, 1]^2 \rightarrow \mathbb{R}^2$  pushes the uniform measure  $du$  on  $[0, 1]^2$  to  $dF$  on  $\mathbb{R}^2$ .

Thus,  $(F_1^{[-1]}(\cdot), F_{X_2}^{[-1]}(\cdot))'$  acts like a *multivariate quantile function*  $Q_F(\cdot) : [0, 1]^2 \rightarrow \mathbb{R}^2$ : monotone and  $Q_F \# du = dF$ .

As a final note, observe that for  $d = 1$ , the univariate distribution function  $F$  and its quantile function  $F^{[-1]}$  (of a real-valued random variable  $X$ ) are referred to as its *rank* and quantile functions, respectively. If  $F$  is continuous, then  $F(X)$  is uniformly distributed on  $[0, 1]$ . By McCann’s theorem applied to  $d = 1$ ,  $F^{[-1]}$  is the (a.e.) unique

monotone map from  $[0, 1]$  to  $\mathbb{R}$  such that  $F^{[-1]} \# du = dF$ , and if, in addition,  $F$  has a finite second moment, then

$$F^{[-1]} = \operatorname{argmin}\{E(U - T(U))^2 : T \# du = dF\}$$

where  $U$  is the random variable uniformly distributed on  $[0, 1]$ .

McCann's theorem for  $d \geq 1$  affirms the existence and uniqueness of a "multivariate" quantile function  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$ , monotone and  $Q_F \# d\mathbf{u} = dF$ , and if, in addition,  $F$  has a finite second moment, then

$$Q_F = \operatorname{argmin}\{E\|U - T(U)\|^2 : T \# d\mathbf{u} = dF\}$$

### 3 Characterization of Univariate Quantile Functions

The notion of (univariate) quantiles is useful in various statistical analyses, mainly in *univariate* quantile regression (Koenker and Bassett [5]) which was developed based on characterizations of other quantities (for computation purposes).

The application of univariate quantile functions to simulations turns out to have a deeper effect.

Recall that a real-valued random variable  $X$  is a measurable map from  $\Omega \rightarrow \mathbb{R}$ , where its source of uncertainty is the "background" probability space  $(\Omega, \mathcal{A}, P)$  and its range space is the measurable space  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , and where  $X$  represents a measure-preserving map transporting the probability measure  $P$ , from its source of uncertainty, to its "law"  $P_X$  on its observation space  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , i.e.  $P \circ X^{-1} = P_X$ . We also denote the law of  $X$  as  $P_X = dF$ , where  $dF((-\infty, x]) = F(x)$ .

Is there some other *concrete and equivalent* source of uncertainty that can replace  $(\Omega, \mathcal{A}, P)$  and a measure-preserving map  $T(\cdot)$  from it to  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ ?

We are asking for another representation of  $X$ . The following is well known in simulations. If  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is the univariate quantile function of  $X$  (or of  $F$ ), and  $U$  is a random variable uniformly distributed on  $[0, 1]$ , then the random variable  $F^{[-1]} \circ U : \Omega \rightarrow \mathbb{R}$  has the same distribution  $F$ , i.e.,  $X \stackrel{D}{=} F^{[-1]} \circ U$ . Thus, if we know  $F$ , we can simulate  $X$ , i.e., obtaining simulated data from  $X$ : pick a random number  $u$  in  $[0, 1]$ , then  $F^{[-1]}(u) = x$  is an outcome from  $X$ . Note that this "concrete" specification of  $X$  (with its source on uncertainty being  $([0, 1], \mathcal{B}([0, 1]), du)$ ) is used for simulation purpose.

*Proof of  $X \stackrel{D}{=} F^{[-1]} \circ U$ .* Let's clarify first the following. If  $Y = T(X)$ , then

$$P(Y \in A) = P(T(X) \in A) = P(X \in T^{-1}(A))$$

so that  $P_Y(A) = P_X(T^{-1}(A))$ . Thus,  $X \stackrel{D}{=} F^{[-1]} \circ U$  means, for any  $A \in \mathcal{B}(\mathbb{R})$ , we have  $dF(A) = du((F^{[-1]})^{-1}(A))$ , where  $du$  denotes the probability measure of  $U$ .

As stated earlier, for probability measures on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , it suffices to consider  $A$  of the form  $A = (-\infty, x]$ . We have

$$(F^{[-1]})^{-1}((-\infty, x]) = \{u \in [0, 1] : F^{[-1]}(u) \in (-\infty, x]\} =$$

$$\{u \in [0, 1] : F^{[-1]}(u) \leq x\} = \{u \in [0, 1] : F(x) \geq u\}$$

since

$$F^{[-1]}(u) \leq x \iff F(x) \geq u$$

therefore,

$$du\{(F^{[-1]})^{-1}((-\infty, x])\} = du\{\{u \in [0, 1] : F(x) \geq u\}\} =$$

$$F(x) = dF((-\infty, x])$$

Thus, we have the “concrete” probability space  $([0, 1], \mathcal{B}([0, 1]), du)$ , replacing the abstract  $(\Omega, \mathcal{A}, P)$ , and the *polar factorization*  $X = F^{[-1]}(U)$  (by analogy of polar factorization of complex numbers, or of matrices) which requires the quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$ , which is a measure-preserving map, pushing the probability measure  $du$  on  $[0, 1]$  to the probability measure  $dF$  on  $\mathbb{R}$ . (in symbol  $dF = (F^{[-1]})\#du$ ).

**Remark.** Following the standard notations in Optimal Transport Theory, when a map  $T$  is a push forward for a probability  $\mu$  to a probability measure  $\nu$ , i.e.,  $\nu = \mu T^{-1}$ , we write  $\nu = T\#\mu$ . Thus,  $X \stackrel{D}{=} F^{[-1]}(U)$  means  $dF = F^{[-1]}\#du$ .

The univariate quantile function  $F^{[-1]}$  satisfies two properties:

- (1)  $F^{[-1]}$  is monotone (non decreasing), and hence the derivative of some convex function,
- (2)  $dF = (F^{[-1]})\#du$ : it pushes  $du$  forward to  $dF$ .

These properties are well known, but what is “new” is that they characterize univariate quantile functions, in the sense that they are obtained from an “abstract” setting, without evoking the total order of the underlying space  $\mathbb{R}$  in the explicit definition of  $F^{[-1]}$ .

Specifically, there is only one map  $T(\cdot) : [0, 1] \rightarrow \mathbb{R}$  satisfying these two conditions. In other words, if a map  $T(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is monotone (non decreasing) and  $dF = T\#du$ , then it is  $F^{[-1]}(\cdot)$ .

Of course, that remains to be proved. But before that, let’s announce what we are going to proceed. Once we prove this characterization of  $F^{[-1]}$ , we can use it to generalize to *multivariate quantile functions of random vectors*, without bother about the lack of a total order relation on  $\mathbb{R}^d$  when  $d > 1$ , thanks to *Optimal Transport Theory* (it is precisely because of OT that Econometricians discover the above characterization of univariate quantile function for generalization to higher dimensions which is so needed in applications, but for so long, no such generalization is available).

Specifically, the two characteristic properties of the univariate quantile function  $F^{[-1]}$  can be addressed on  $\mathbb{R}^d$  when  $F^{[-1]}$  is replaced by a map  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$  which is monotone (non decreasing) as: for any  $x, y \in [0, 1]^d$ ,

$$\langle Q_F(x) - Q_F(y), x - y \rangle \geq 0$$

where  $\langle \cdot, \cdot \rangle$  is the scalar product on  $\mathbb{R}^d$ . Clearly, the property (2) is meaningful on any probability spaces.

Note that the property (1) is very important! even, usually we do not emphasize it. Being a monotone non decreasing function,  $F^{[-1]}$  is qualified as the derivative of a some convex function. For example, if  $F^{[-1]}(u) = u + 1$ , then it is the derivative of the convex function  $\frac{1}{2}(u + 1)^2$ . Note that, for a convex function  $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ , its gradient (vector of partial derivatives)  $\nabla\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is monotone non decreasing in the above sense.

So what we will do next in this Section is to show the following. Consider the probability spaces  $([0, 1], \mathcal{B}([0, 1]), du)$ , and  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), dF)$ . While we know that  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is a map having two properties (1), (2) above, we need to show two more things, namely its uniqueness, and optimality. Why? Well, our purpose is to characterize  $F^{[-1]}$  in a setting (which will be Optimal Transport/OT) suitable for generalizing to higher dimensions.

Without exaggeration, it can be said that, like Copulae, OT theory will invade statistics of this 21st century!

*Uniqueness of  $F^{[-1]}$ .* Suppose  $T$  is a monotone non decreasing map and  $T\#du = dF$ . We are going to show that  $T = F^{[-1]}$  so that  $F^{[-1]}$  is unique.

*Proof.* By monotonicity of  $T$ , we have

$$(-\infty, x] \subseteq T^{-1}((-\infty, T(x)])$$

so that

$$F_{du}(x) = du(-\infty, x] \leq du\{T^{-1}((-\infty, T(x)))\} = dF(-\theta, T(x)) = F(T(x))$$

and  $T(x) \geq F^{[-1]}(x)$ .

Suppose the inequality is strict. This means that there exists  $\varepsilon_o > 0$  such that  $F(T(x) - \varepsilon) \geq F_{du}(x)$  for every  $\varepsilon \in [0, \varepsilon_o]$ . Also, since  $T^{-1}((-\infty, T(x) - \varepsilon)) \subseteq (-\infty, x)$ , we have  $F(T(x) - \varepsilon) < F_{du}(x)$ . Thus,  $F(T(x) - \varepsilon) = F_{du}(x)$  for any  $\varepsilon \in [0, \varepsilon_o]$ . Note that  $F(T(x) - \varepsilon)$  is the value of  $F$  which  $F$  takes on an interval where it is constant. But these intervals are a countable quantity, so that the values  $y_j$  of  $F$  on these intervals are also countable. Therefore, the points  $x$  where  $T(x) > F^{[-1]}(x)$  are contained in  $\cup_j \{x : F_{du}(x) = y_j\}$  which is  $du$ -negligible (since  $du$  is atomless). As a consequence,  $T(x) = F^{[-1]}(x)$ ,  $du$ -almost everywhere.

**Remark.** More generally, if  $\mu, \nu$  are Borel probability measures on  $\mathbb{R}$ , with supports  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$ , respectively, with  $\mu$  atomless, there is a unique, monotone non decreasing transport map, namely  $x \rightarrow F_\nu^{[-1]}(F_\mu(x))$ .

*Optimality of  $F^{[-1]}$ .* A transport map (monotone or not)  $T(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is a measure-preserving map, i.e.,  $T\#du = dF$ . By optimality, we mean the following. Let  $c(\cdot, \cdot) :$

$[0, 1] \times \mathbb{R} \rightarrow \mathbb{R}^+$  be a “cost” function (of transporting elements of  $[0, 1]$  to elements of  $\mathbb{R}$ ). A transport map  $T^*$  is optimal, with respect to  $c$  if

$$T^* = \arg \min \left\{ \int_0^1 c(u, T(u)) du : T : T\#du = dF \right\}$$

If the cost function is of the form  $c(u, x) = h(u - x)$  with  $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  strictly convex (such as  $h(y) = y^2$ ), then, independent of  $c$ ,  $F^{[-1]}$  is optimal, i.e., we got an explicit formula for the unique monotone transport map, in this one-dimensional case. We will illustrate this via an example here. With a bit of Optimal Transport theory in the next Section, we will provide a general theorem in higher dimensions, together with a dual formulation for the computation of the optimal solution.

Let  $([0, 1], \mu = du)$ , and  $([1, 2], \nu = dv)$ , where  $dv$  is uniform on  $[1, 2]$  ( $dv = dF$ ), and  $F^{[-1]}(\cdot) = [0, 1] \rightarrow \mathbb{R} : F^{[-1]}(x) = x + 1$ . We will check that it is the unique monotone and optimal map. Clearly, it is monotone (and is the derivative of some convex function, e.g.,  $\frac{1}{2}(1+x)^2$ ). That  $F^{[-1]}\#du = dv$  because  $X \stackrel{D}{=} F^{[-1]}(U)$  where  $U \simeq du$ , and  $X \simeq dv$ . By the above proof of uniqueness, it is the only monotone map pushing  $du$  on  $[0, 1]$  to  $dv$  on  $[1, 2]$ .

It remains to show that it is optimal with respect to convex cost function, such as  $c(\cdot, \cdot) : [0, 1] \times [1, 2] \rightarrow \mathbb{R}^+$ ,  $c(u, v) = h(u - v)$ , with  $h(\cdot)$  convex, e.g.,  $h(x) = x^2$ .

Let  $T(\cdot) : [0, 1] \rightarrow [1, 2]$  be a transport map, i.e.,  $T\#du = dv$ , monotone or not. The total cost of  $T$  is

$$C(T) = \int_0^1 (u - T(u))^2 du$$

We have, using Jensen’s inequality ( $h(EX) \leq Eh(X)$ ):

$$\begin{aligned} C(T) &= \int_0^1 h(T(x) - x) dx \geq h \left[ \int_0^1 (T(x) - x) dx \right] = \\ &= h \left[ \int_0^1 T(x) dx - \int_0^1 x dx \right] = \end{aligned}$$

Note that the following are non-monotone transport maps:  $T(x) = 2 - x$ , and

$$S(x) = \begin{cases} x + \frac{3}{2} & \text{for } x \in [0, \frac{1}{2}] \\ 2 - x & \text{for } x \in [\frac{1}{2}, 1] \end{cases}$$

with

$$C(T) = \int_0^1 (x - T_2(x))^2 dx = \int_0^1 (2x - 2)^2 dx = \frac{4}{3}$$

$$C(S) = \int_0^1 (x - T_3(x))^2 dx = \int_0^{\frac{1}{2}} \left(\frac{3}{2}\right)^2 dx + \int_{\frac{1}{2}}^1 (2x - 2)^2 dx = \frac{31}{24}$$

whereas

$$C(F^{[-1]}) = \int_0^1 (x - T_1(x))^2 dx = 1$$

which is the smallest.

Note that, in the above calculations, we only use the fact that  $h(\cdot)$  is (strictly) convex, but not its specific form. Thus, in fact,  $F^{[-1]}$  is optimal with respect to any convex loss.

## 4 Optimal Transport and Multivariate Quantiles

We elaborate a bit on the theory of Optimal Transport from which to derive multivariate quantile functions.

As the polar factorization of a real-valued random variable is  $X \stackrel{D}{=} F^{[-1]}(U)$ , we are looking for the polar factorization of a random vector  $\mathbf{X} \stackrel{D}{=} Q(U)$ .

We are interested in the question: What could be the counterpart of a univariate quantile function  $F^{[-1]}$  in higher dimensions, i.e., for a multivariate distribution function  $F$  on  $\mathbb{R}^d$ , with  $d > 1$ ? The lack of a total order on  $\mathbb{R}^d$  seems responsible for unsuccessful attempts in the past.

We are interested in quantile functions of distribution functions for a variety of reasons. We know very well what is the quantile function  $F^{[-1]}$  explicitly of a univariate distribution function  $F$  of a random variable  $X$ , for arbitrary distribution function, heavy-tailed or not.

The characterization of  $F^{[-1]}$  in Sect. 3 serves as a prototype for a generalization to higher dimensions. Thus, first, we call upon McCann's theorem to have the existence of a unique measure-preserving map  $T : [0, 1]^d \rightarrow \mathbb{R}^d$ ,  $T\#du = dF$ , where  $du$  is the uniform probability measure on  $[0, 1]^d$ , and  $F$  is an arbitrary multivariate distribution function on  $\mathbb{R}^d$ , with  $T$  being monotone. Then, we rely upon Brenier's theorem to emphasize that such  $T$  in McCann's theorem is "optimal" in Monge's problem (MP) which, in fact, also optimal in Kantorovich extended problem (KP). While (KP) is "solvable", we need a dual formulation to get the solution, via linear programming, and finally obtain a computable form of our desired transport map which will be our multivariate quantile function for the multivariate distribution function  $F$  on  $\mathbb{R}^d$ .

In one dimension, the quantile function  $F^{[-1]}$  of the univariate distribution function  $F$  (of a real-valued random variable  $X$ ) is the unique monotone map from  $[0, 1]$  to  $\mathbb{R}$  such that  $X \stackrel{D}{=} F^{[-1]}(U)$  (a polar factorization of  $X$ ), where  $U$  is the random variable uniformly distributed on  $[0, 1]$ , i.e., with  $F_U(u) = u$ , or with probability measure  $du$  on  $[0, 1]$ , or equivalently  $dF = du \circ (F^{[-1]})^{-1}$ , in symbol.  $F^{[-1]}\#du = dF$ .

Quick question: Is there a polar factorization for  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ , with  $d > 1$ ? The answer is yes!

*McCann's Theorem.* There exists a unique ( $du$ -a.e., where  $du$  is the uniform probability measure on  $[0, 1]^d$ ) measurable map  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$  which is the gradient of some convex function  $\varphi(\cdot) : [0, 1]^d \rightarrow \mathbb{R}$  (hence monotone) and such that  $Q_F\#du = dF$  (i.e.,  $du \circ Q_F^{-1}(\cdot) = dF(\cdot)$ ).

Thus, if we let  $X$  be the random vector with multivariate distribution function  $F$  on  $\mathbb{R}^d$ , and  $U$  being the random vector uniformly distributed on the unit cube  $[0, 1]^d$ , then we have  $X \stackrel{D}{=} Q_F(U)$ .

Note that the *multivariate quantile function*  $Q_F$  of  $F$  on  $\mathbb{R}^d$  exists for any distribution functions  $F$  (just like in dimension 1 where  $F^{[-1]}$  is defined regardless whether  $dF$  has finite moments or not). In dimension 1,  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is monotone and  $F^{[-1]} \# du = dF$ , it is  $\nabla\varphi$  in view of the uniqueness in McCann’s theorem.

When  $dF$  has finite second moment,  $\nabla\varphi = F^{[-1]}$  in McCann’s theorem is “optimal”, with respect to square loss function, in the following Monge’s problem (MP):

$$\nabla\varphi = \arg \min \left\{ \int_0^1 (x - T(x))^2 dx : T \# du = dF \right\}$$

as we have seen in Sect. 3. In fact,  $F^{[-1]}$  can be determined as  $\nabla\varphi$  in the MP above. In fact, this situation is general (by Brenier’s theorem).

If we are just interested in the existence of vector quantiles (i.e., quantile functions of random vectors) then McCann’s theorem is enough. However, if we want to use vector quantiles to conduct, say, multivariate quantile regression, or to define multivariate (financial) risk measures, then their existence is not enough! We need to determine them explicitly for applications.

For dimension  $d > 1$ , the situation is not simple (!) as the Monge’s minimization is somewhat intractable because its objective function is not linear in  $T$ , and the constraint set  $\{T : T \# du = dF\}$  is not convex. We need to avoid these difficulties by embedding (MP) into the Kantorovich problem (KP) to use linear programming in its dual formulation. Such a program will help us to “compute” multivariate quantile functions. Thus, we need to evoke a bit of Optimal Transport (OT) theory.

Roughly speaking, observe that if  $T$  is in the constraint set of (MP), then  $du \circ (I, T)^{-1}$  is a joint probability measure on  $[0, 1] \times \mathbb{R}$  having  $du, dF$  as marginal measures, we consider the (larger) convex constraint set  $\Pi(du, dF)$  of all joint measures with  $du, dF$  as marginals, and the linear objective function (in  $\pi \in \Pi(du, dF)$ )

$$\pi \rightarrow \int_{[0,1] \times \mathbb{R}} c(x, y) d\pi(x, y)$$

and address the Kantorovich problem (KP)

$$\min \left\{ \int_{[0,1] \times \mathbb{R}} c(x, y) d\pi(x, y) : \pi \in \Pi(du, dF) \right\}$$

which is “tractable”, thanks to duality. If the (KP) has a solution of the form  $\pi_T = du \circ (I, T)^{-1}$  then  $T$  will be a solution of the (MP).

To complete our agenda description, here is what we will proceed. The (MP) is enlarged to (KP) in view of

$$\{T : T \# du = dF\} \subseteq \Pi(du, dF)$$

by the identification of  $T$  with  $\pi_T = du \circ (I, T)^{-1}$ . While the (KP) is linear under convex constraint set, its constraint set is not expressed as inequalities (in infinitely

dimensional form). Thus, we need to use duality, i.e., relating the “inf” problem of (KP) to a “sup” problem whose constraint set is expressed as inequalities, and then using linear programming to solve it, noting that Kantorovich is the inventor of linear programming (for solving Monge’s original problem).

While we seek a candidate for a multivariate quantile function of a distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ , generalizing the well-known univariate quantile function  $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$  which is characterized as the unique monotone map pushing the uniform probability measure  $du$  on  $[0, 1]$  to  $dF$ , namely a unique map  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$ , monotone and pushing the uniform probability on  $[0, 1]^d$  to the probability measure  $dF$  on  $\mathbb{R}^d$ , we have McCann’s theorem affirming the existence and uniqueness of a such candidate, we still need to obtain it constructively for applications.

The roads leading to them are as follows. First, we extend (MP) to (KP) to make sure that there are solutions for (KP) which came from solutions of (MP), i.e., of the form  $\gamma_T = (I, T)\#du$ . For the strict convex loss  $c(x, y) = h(x - y)$  with  $h(t) = \frac{t^2}{2}$  or  $t^2$ , it turns out that there exists uniquely ( $du$ -a.e.) an optimal  $\gamma_T$  for (KP), for which, the associated  $T$  is optimal for (MP). How to determine that  $T$ ? (which will be our desired  $Q_F$ ). We need results from duality. The unique optimal pair  $(\varphi, \varphi^*)$  of the dual problem, where  $\varphi^*$  is the  $c$ -transform of  $\varphi$ , is related to  $T$  as  $T(x) = x - (\nabla h)^{-1}(\nabla \varphi(x))$  (which is the gradient of the convex function  $x \rightarrow \frac{x^2}{2} - \varphi(x)$ ). Thus,  $T(\cdot)$  is determined once we can determine  $\varphi$  in the dual problem.

*Example.* Let  $\mu = dF$  and  $\nu = dG$  on  $\mathbb{R}$ , and  $c(x, y) = (x - y)^2$ . Then  $\pi^* = (F^{[-1]}, G^{[-1]})\#du$ , for  $du$  uniform on  $[0, 1]$ , is optimal for (KP). Thus, for  $dF = du$ , we have  $\pi^* = (I, G^{[-1]})\#du$ , i.e.,  $\pi^* = \gamma_T = (I, T)\#du$ , with  $T = G^{[-1]}$  optimal for (MP).

As we have elaborated in Sect. 3, the concept of a “transport map” appeared already from the beginning of probability theory. Indeed, if  $X$  is a (real-valued) random variable, defined on a probability space  $(\Omega, \mathcal{A}, P)$ , then  $X$  acts like a map from  $\Omega$  to the real line  $\mathbb{R}$  (the observed values of  $X$  are “outcomes” or results from what happened in  $\Omega$ ), transporting the probability measure  $P$  on  $\Omega$  to its law  $P_X$  on  $\mathbb{R}$ , in the sense that  $P_X = PX^{-1}$ , in symbol  $X\#P = P_X$ .

In fact we have a more concrete transport map which is the (univariate) quantile function  $F^{[-1]}$  which is a map transporting the uniform probability measure  $du$  on  $[0, 1]$  to  $P_X$  (or  $dF$ ) on  $\mathbb{R}$ , i.e.,  $F^{[-1]}\#du = dF$  (the polar factorization of  $X$ ). In both settings, we have a map which preserves probabilities. In other words,  $X : \Omega \rightarrow \mathbb{R}$ , and  $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$  are measure-preserving maps.

What is *optimal* transport problem? In 1781 Gaspard Monge considered the following problem. Let  $(\mathcal{X}, \mu), (\mathcal{Y}, \nu)$ , with  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ , say, be two (Borel) probability spaces, and  $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be a cost function (of transporting elements of  $\mathcal{X}$  to elements of  $\mathcal{Y}$ ). Find the best (optimal) preserving map  $T^*$  which transports  $\mu$  (mass distribution) to  $\nu$ , i.e.,  $T^*\#\mu = \nu$ , in the sense of minimizing the total transport cost, i.e.,

$$T^* = \inf \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) : T : T\#\mu = \nu \right\}$$

In our analysis, we can take  $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ ,  $\nu = dF$  the probability measure associated with the multivariate distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ , and  $\mu = du$ , the



(non atomic) uniform probability measure on  $[0, 1]^d$  which is considered as on  $\mathbb{R}^d$ , as follows.

For  $d = 1$ , the distribution of the variable  $U$ , uniformly distributed on  $[0, 1]$  has the distribution  $F_U(\cdot)$ , and its associated probability measure  $dF_U(-\infty, x] = F_U(x) = x$ , for  $x \in [0, 1]$ .

For  $d > 1$ , if  $U$  is uniformly distributed on  $[0, 1]^d$ , then its distribution function  $F_U(\cdot) : [0, 1]^d \rightarrow [0, 1]$  is

$$F_U(u_1, u_2, \dots, u_d) = \prod_{j=1}^d u_j \quad \text{for } (u_1, u_2, \dots, u_d) \in [0, 1]^d$$

i.e.,  $dF_U$  is the product measure with uniform marginals on  $[0, 1]$ , a special  $d$ -copula.

We have seen an example, in Sect. 3, of this problem. Specifically, if  $(\mathcal{X}, \mu), (\mathcal{Y}, \nu)$  are  $([0, 1], du)$  (a nonatomic probability measure with finite second moment), and  $(\mathbb{R}, dF)$  (an arbitrary probability measure), respectively, then the Monge’s optimal transport map with respect to a convex loss function, e.g.,  $c(x, y) = (x - y)^2$ , is  $F^{[-1]}$ .

Moreover, the optimal transport map  $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$  is an unique monotone non decreasing map, qualifying as the derivative of some convex function on  $[0, 1]$ .

Let’s reexamine this example again. We know in advance that the Monge’s solution must be  $F^{[-1]}$  since the quantile function  $F^{[-1]}(x) = x + 1$  (where  $F(\cdot)$  is the uniform distribution function on  $[1, 2]$ ) satisfies the two basic properties of a monotone optimal map, and in view of the uniqueness of such a map. But can we actually get that explicit optimal map without knowing the notion of univariate quantile functions? and “define”  $F^{[-1]}$  as such?

Answering the above question opens the door for defining and determining multivariate quantile functions.

*From Monge to Kantorovich.* Since we are only interested in defining multivariate quantile functions, we will consider only two specific probability spaces  $([0, 1]^d, \mathcal{B}([0, 1]^d), du)$  and  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), dF)$ , where  $du$  is the uniform probability measure on  $[0, 1]^d$ , with uniform distribution function  $F_U$  on  $[0, 1]^d$ , and  $dF$  is the probability measure associated with the multivariate distribution function  $F$  on  $\mathbb{R}^d$ .

In terms of random variables, we refer to  $U$  as the uniform random vector with distribution  $F_U$ , and  $X$  as the random vector with distribution function  $F$ .

We will not need to consider the general theory of Optimal Transport (OT).

For dimension  $d = 1$ , we have the explicit form of the univariate quantile function  $F^{[-1]}$  (which is the “solution” of Monge’s problem for convex loss functions), and, we will have existence and uniqueness of its counterpart in any dimension  $d > 1$ , without evoking OT. However, for computations of Monge’s solutions in higher dimensions, we need to address them in the setting of OT, in order to use linear programming in a dual formulation. Thus, we will mention a bit of OT which is beneficial in larger contexts.

The Monge’s problem can be extended to a more general formulation (due to Kantorovich) as follows. Let  $(\mathcal{X}, \mu), (\mathcal{Y}, \nu)$  be Borel probability spaces with  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ .

Let  $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be a cost function. Then the solution problem to the Monge’s problem is an optimal transport map  $T^*(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ , i.e.,

$$T^* = \arg \min \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) : T : T\#\mu = \nu \right\}$$

In general, Monge's problem might not even have solutions. And when it does have solutions, it is not easy to compute them, because the objective function is not linear, and the constraint set is not convex.

The Kantorovich's reformulation avoids these difficulties.

Kantorovich's formulation is based on the idea of enlarging Monge's problem so that, first of all, it always has solutions. This is somewhat similar to the introduction of complex numbers, or more closely to von Neumann's mixed strategies in game theory (extending pure (determinist) strategies to random strategies).

Observe that, if  $T$  is a transport map, i.e.,  $T\#\mu = \nu$  ( $\mu T^{-1}(\cdot) = \nu(\cdot)$ ), then, denoting by  $I$  the identity function on  $\mathcal{X}$ ,  $\gamma_T = (I, T)\#\mu$  is the probability measure on  $\mathcal{X} \times \mathcal{Y}$  admitting  $\mu, \nu$  as marginal measures.

**Remark.** The map  $(I, T) : \mathcal{X} \rightarrow \mathbb{R}^2$ , is defined as  $(I, T)(x) = (x, T(x)) \in \mathbb{R}^2$ .

The joint measure  $\gamma_T = (I, T)\#\mu$  is characterized as: for any  $f(\cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ , we have

$$\int_{\mathcal{X} \times \mathcal{Y}} f(x, y) d\gamma_T(x, y) = \int_{\mathcal{X}} f(x, T(x)) d\mu(x)$$

*Proof.* Use "standard argument of measure theory", starting out with  $f$  being an indicator function, i.e.,  $f(x, y) = 1_{A \times B}(x, y)$ . Then we have

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) d\gamma_T(x, y) &= \int_{\mathcal{X} \times \mathcal{Y}} 1_{A \times B}(x, y) d\gamma_T(x, y) = \int_{A \times B} d\gamma_T(x, y) = \\ d\mu(I, T)^{-1}(A \times B) &= d\mu(A \cap T^{-1}(B)) = \int_{\mathcal{X}} 1_{A \times B}(x, T(x)) d\mu(x) \end{aligned}$$

Indeed,

$$(I, T)\#\mu(A \times \mathcal{Y}) = \mu(I, T)^{-1}(A \times \mathcal{Y}) = \mu\{x \in \mathcal{X} : (I, T)(x) \in A \times \mathcal{Y}\} =$$

$$\mu\{x \in \mathcal{X} : (x, T(x)) \in A \times \mathcal{Y}\} = \mu\{x \in \mathcal{X} : x \in A\} = \mu(A)$$

and

$$(I, T)\#\mu(\mathcal{X} \times B) = \mu(I, T)^{-1}(\mathcal{X} \times B) = \mu\{x \in \mathcal{X} : (I, T)(x) \in \mathcal{X} \times B\} =$$

$$\mu\{x \in \mathcal{X} : (x, T(x)) \in \mathcal{X} \times B\} = \mu\{x \in \mathcal{X} : x \in \mathcal{X}, T(x) \in B\} =$$

$$\mu\{x \in \mathcal{X} : x \in \mathcal{X}, T(x) \in B\} = \mu\{x : x \in T^{-1}(B)\} = \nu(B)$$

Thus, if  $c(.,.) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  be a given cost function, we have

$$V_c(\gamma_T) = \int_{\mathcal{X} \times \mathcal{Y}} c(x,y)d\gamma_T(x,y) = \int_{\mathcal{X}} c(x,T(x))d\mu(x) = V_c(T)$$

where  $V_c$  denote the value of the transport plan  $\gamma_T$  and of the transport map  $T$ , with respect to  $c$ .

Thus, Monge’s transport maps are special cases of transport “plans” (i.e., joint probability measures on  $\mathcal{X} \times \mathcal{Y}$  having  $\mu, \nu$  as marginals).

Thus, if we denote by  $\Pi(\mu, \nu)$  the set of joint probability measures on  $\mathcal{X} \times \mathcal{Y}$  having  $\mu, \nu$  as marginals, then we enlarge the setting of Monge’s problem (MP) in which  $\Pi(\mu, \nu)$  is the solution set for the Kantorovich problem (KP):

$$\min\left\{\int_{\mathcal{X} \times \mathcal{Y}} c(x,y)d\pi(x,y) : \pi \in \Pi(\mu, \nu)\right\}$$

which always has solutions since  $\Pi(\mu, \nu) \neq \emptyset$  (the product measure  $\mu \otimes \nu \in \Pi(\mu, \nu)$ ).

Note that if  $\gamma_T$  is a solution for (KP), then  $T$  is solution for (MP).

For example, if we take  $\mu = du$ , and  $\nu = dF$  in dimension 1 (i.e., on  $\mathbb{R}$ ), then  $\gamma = du \circ (I, F^{[-1]})^{-1} \in \Pi(\mu, \nu)$ , noting that the identity  $I$  on  $[0, 1]$  is the quantile function of the uniform distribution.

Note that the dependence structure of the random variables  $U$  (uniform on  $[0, 1]$  with distribution function  $F_U(u) = u$ ) and  $X$  (with distribution function  $F$ ) in the polar factorization  $X = F^{[-1]}(U)$  is that  $U$  and  $X$  are *comonotone*, i.e., they go up or down together (the subset  $\{U(\omega), X(\omega) : \omega \in \Omega\}$  is totally ordered in  $\mathbb{R}^2$ , which is the subset  $\{(x, y) \in \mathbb{R}^2\}$  such that, for any,  $(x, y), (x', y')$  in it, we have  $\langle x - x', y - y' \rangle \geq 0$ ). According the Sklar’s theorem,  $U$  and  $X$  are comonotone if and only if the *copula of their dependence structure* is  $\mathcal{C}(u, \nu) = u \wedge \nu$ . This can be seen as follows. The joint measure of  $du, dF$  is  $\gamma = du \circ (I, F^{[-1]})^{-1}$ , so that the associated joint distribution function of  $(U, X)$  is

$$H(a, b) = P(U \leq a, X \leq b) = \gamma((-\infty, a] \times (-\infty, b])$$

then

$$H(a, b) = dH((-\infty, a] \times (-\infty, b]) = du(I, F^{[-1]})^{-1}((-\infty, a] \times (-\infty, b]) =$$

$$du\{u : u \leq a, F^{[-1]}(u) \leq b\} = du\{u \leq a, F(b) \geq u\} =$$

$$du\{u : u \leq a \wedge F(b)\} = a \wedge F(b) = F_U(a) \wedge F(b)$$

so that  $U$  and  $X$  are comonotone, or, by abuse of language, their joint probability measure  $\gamma = du \circ (I, F^{[-1]})^{-1}$  is comonotone.

Therefore, if  $T$  is a monotone transport map then its corresponding transport plan  $\gamma_T = du \circ (I, T)^{-1}$  is comonotone.

**Remark.** In fact, the above can be extended to two variables, namely the joint measure  $\gamma = (F_Y^{[-1]}, F_X^{[-1]})\#du \in \Pi(dF_Y, dF_X)$  is comonotone.

Indeed, let's verify first that  $\gamma = (F_Y^{[-1]}, F_X^{[-1]})\#du \in \Pi(dF_Y, dF_X)$ . We have

$$\gamma(A \times \mathbb{R}) = du((F_Y^{[-1]}, F_X^{[-1]})^{-1}(A \times \mathbb{R}) =$$

$$du\{u \in [0, 1] : F_Y^{[-1]}(u) \in A, F_X^{[-1]}(u) \in \mathbb{R}\} =$$

$$du\{u \in [0, 1] : F_Y^{[-1]}(u) \in A\} = F_Y^{[-1]}\#du(A) = dF_Y(A)$$

Similarly,

$$\gamma(\mathbb{R} \times B) = dF_X(B)$$

Next, we have

$$H(a, b) = P(Y \leq a, X \leq b) = \gamma((-\infty, a] \times (-\infty, b]) =$$

$$du\{u \in [0, 1] : F_Y^{[-1]}(u) \leq a, F_X^{[-1]}(u) \leq b\} =$$

$$du\{u \in [0, 1] : F_Y(a) \geq u, F_X(b) \geq u\} =$$

$$du\{u \in [0, 1] : u \leq F_Y(a) \wedge F_X(b)\} = F_Y(a) \wedge F_X(b)$$

**Remark.** The space  $\Pi(du, dF)$  can be specified as follows. For  $F(\cdot)$  continuous, each  $\gamma \in \Pi(du, dF)$  is indexed by a (binary) copula  $\mathcal{C}$ , say  $\gamma_{\mathcal{C}}$ , since the joint distribution function  $H_{\mathcal{C}}$  of  $\gamma_{\mathcal{C}}$ , i.e.,  $\gamma_{\mathcal{C}} = dH_{\mathcal{C}}$ , is  $H_{\mathcal{C}}(u, x) = \mathcal{C}(u, F(x))$ . Thus, each copula  $\mathcal{C}$  determines a joint measure  $\gamma_{\mathcal{C}} \in \Pi(du, dF)$ . In particular, for  $\mathcal{C}(u, v) = u \wedge v$ , we get  $\gamma_{\mathcal{C}} = du \circ (I, F^{[-1]})^{-1}$  (corresponding to the extremal copula). If  $F(\cdot)$  is not continuous (e.g., it's an empirical distribution) we must include sub-copulas.

For  $\mu, \nu \in \mathcal{P}(\mathbb{R})$ , set of Borel probability measures on  $\mathbb{R}$ , a joint measure  $\gamma \in \Pi(\mu, \nu)$  is such that  $\gamma(A \times \mathbb{R}) = \mu(A)$ ,  $\gamma(\mathbb{R} \times B) = \nu(B)$ . For example, let  $\mathcal{C}$  be a bivariate copula, then  $\gamma_{\mathcal{C}} \in \Pi(\mu, \nu)$  is determined by  $\gamma_{\mathcal{C}} = dH_{\mathcal{C}}$ , where  $H_{\mathcal{C}}(x, y) = \mathcal{C}(F_{\mu}(x), F_{\nu}(y))$ . For  $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ , with  $d > 1$ , this procedure will need the generalization of copulas to *vector copulas*.

The above characteristics of the univariate quantile function  $F^{[-1]}$  is considered as its definition, i.e., let  $F$  be an arbitrary (univariate) distribution function, then its (univariate) quantile function is the unique monotone non decreasing optimal transport map between  $([0, 1], du)$  and  $(\mathbb{R}, dF)$  with respect to a convex loss function.

What we have in mind is this. Let  $(\mathcal{X}, \mu), (\mathcal{Y}, \nu)$ , with  $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ ,  $d > 1$ , where  $(\mathcal{X}, \mu) = ([0, 1]^d, du)$ ,  $(\mathcal{Y}, \nu) = (\mathbb{R}^d, dF)$ , and  $c(x, y) = \|x - y\|^2$ . If there exists an unique gradient  $\nabla\phi$  (of some convex function (not unique)  $\phi : [0, 1]^d \rightarrow \mathbb{R}$ ) which is

the optimal transport, then  $\nabla\varphi$  is defined as the multivariate quantile function of  $F$ , noting that the gradient  $\nabla\varphi$  is monotone non decreasing as the generalization of the same concept in one dimension, i.e., for any  $x, y \in \mathbb{R}^d$ , we have

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y \rangle \geq 0$$

where  $\langle \cdot, \cdot \rangle$  denotes the scalar product on  $\mathbb{R}^d$ .

Of course such a result is only an existence result. We need to find ways to compute it, at least for applications!

*Notes on Convex Functions.* A function  $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  is said to be convex if for any  $x, y \in \mathbb{R}^d$  and  $t \in [0, 1]$  we have

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

it is strictly convex if the above inequality is strict.

A convex function is a.e. differentiable. In dimension 1, the graph of a convex function lies above any tangent to it, and hence its derivative is monotone non decreasing. For  $d > 1$ , the whole graph of  $f(\cdot)$  lies above its tangent hyperplane at any  $xc$  where it is differentiable, so as a consequence, its gradient is monotone in the above sense.

The Kantorovich’s reformulation (of Monge’s problem) is this. Find an optimal transport plan, i.e., a joint measure  $\pi^* \in \Pi(\mu, \nu)$  such that

$$\pi^* = \arg \min \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}$$

Now the problem seems solvable since the objective function is linear in  $\pi$ , and the constraint set is convex.

Note that, although, as far as quantile functions are concerned, we are interested in transport maps (not necessary transport plans), we still need to evoke Kantorovich’s formulation in order to compute multivariate quantile functions.

*Duality.* In order to solve the “inf” problem of (KP) we will transform it into a “sup” problem (this procedure is referred to as duality, where the “inf” is the primal problem, and the “sup” is dual problem) where the constraints in the “sup” problem can be expressed as inequalities (In infinite dimensions). The relations between the primal and dual problems will allow us to get solution for the primal problem from the dual problem.

For  $\pi \in \Pi(\mu, \nu)$ , let

$$V(\pi) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y)$$

we are going to relate the (KP)  $P = \inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\}$  to a “sup” problem. For that, first observe that, for suitable function  $\varphi, \psi$  defined on  $\mathcal{X}, \mathcal{Y}$ , respectively, we have

$$\int_{\mathcal{X} \times \mathcal{Y}} [\varphi(x) + \psi(y)] d\pi(x, y) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y)$$

*Proof.* Use “standard argument of measure theory”!

Thus, for  $\varphi, \psi$  such that  $\varphi(x) + \psi(y) \leq c(x, y)$ , for all  $x, y$  we have

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) \leq \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y)$$

Consider

$$D = \sup\left\{ \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) : (\varphi, \psi) : \varphi(\cdot) + \psi(\cdot) \leq c(\cdot, \cdot) \right\}$$

then clearly  $D \leq P$ . In fact,  $D = P$  which is our desired duality. The dual formulation has a linear objective function with inequality constraints (suitable for linear programming).

The Kantorovich duality is this (1942). Let

$$J(\varphi, \psi) = \left\{ \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) \right\}$$

Then

$$P = \inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\} = \sup\{J(\varphi, \psi) : \varphi(\cdot) + \psi(\cdot) \leq c(\cdot, \cdot)\} = D$$

The sup on the right hand side is attained.

We study the duality in the case of quadratic cost  $c(x, y) = \frac{1}{2} \|x - y\|^2$ , when  $\mu$  and  $\nu$  have finite second moments, i.e.,

$$\int_{\mathbb{R}^d} \|x\|^2 d\mu(x) < \infty, \quad \int_{\mathbb{R}^d} \|y\|^2 d\nu(y) < \infty$$

so that

$$V(\pi) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{\|x - y\|^2}{2} d\pi(x, y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{\|x\|^2 + \|y\|^2}{2} d\pi(x, y) < \infty$$

From the duality

$$\inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\} = \sup\{J(\varphi, \psi) : \varphi(\cdot) + \psi(\cdot) \leq c(\cdot, \cdot)\}$$

we get, for the (KP) primal problem: The left hand side admits a minimizer, i.e., there exists  $\pi^* \in \Pi(\mu, \nu)$  such that

$$V(\pi^*) = \inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\}$$

As for the dual problem, here  $\varphi(x) + \psi(y) \leq c(x, y)$  means

$$\varphi(x) + \psi(y) \leq \frac{\|x - y\|^2}{2} = \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - \langle x, y \rangle$$

from it, we have

$$\langle x, y \rangle \leq \left[ \frac{\|x\|^2}{2} - \varphi(x) \right] + \left[ \frac{\|y\|^2}{2} - \psi(y) \right]$$

Let

$$\tilde{\varphi}(x) = \frac{\|x\|^2}{2} - \varphi(x), \quad \tilde{\psi}(y) = \frac{\|y\|^2}{2} - \psi(y)$$

and

$$M = \int_{\mathbb{R}^d} \|x\|^2 / 2 d\mu(x) < \infty + \int_{\mathbb{R}^d} \|y\|^2 / 2 d\nu(y) < \infty$$

we have

$$\inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\} = M - \sup\left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}$$

and

$$\sup\{J(\varphi, \psi) : \varphi(\cdot) + \psi(\cdot) \leq c(\cdot, \cdot)\} = M - \inf\{J(\varphi, \psi) : (\varphi, \psi) \in \tilde{\Theta}\}$$

where

$$\tilde{\Theta} = \{(\varphi, \psi) : \nabla_x \langle x, y \rangle : \varphi(x) + \psi(y) \geq \langle x, y \rangle\}$$

and

$$\langle x, y \rangle \leq \left[ \frac{\|x\|^2}{2} - \varphi(x) \right] + \left[ \frac{\|y\|^2}{2} - \psi(y) \right]$$

becomes

$$\sup\left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\} = \inf\{J(\varphi, \psi) : (\varphi, \psi) \in \tilde{\Theta}\}$$

Let  $\tilde{\varphi}(x) = \left[ \frac{\|x\|^2}{2} - \varphi(x) \right]$ , and  $\tilde{\psi}(y) = \left[ \frac{\|y\|^2}{2} - \psi(y) \right]$ , we have the constraint  $\tilde{\varphi}(x) + \tilde{\psi}(y) \geq \langle x, y \rangle$ .

For simplicity, we just drop the symbol  $\sim$  on the functions  $\varphi, \psi$  from our writing (but not from our mind).

Thus, from  $\varphi(x) + \psi(y) \geq \langle x, y \rangle$ , we have

$$\psi(y) \geq \langle x, y \rangle - \varphi(x) \implies \psi(y) \geq \sup_x [\langle x, y \rangle - \varphi(x)] = \varphi^*(y)$$

so that

$$J(\varphi, \psi) \geq J(\varphi, \varphi^*)$$

We call  $(\varphi, \varphi^*)$  a potential pair. Note that, from

$$\varphi^*(y) = \sup_x [\langle x, y \rangle - \varphi(x)]$$

it follows that  $\varphi(x) + \varphi^*(y) \geq \langle x, y \rangle$ , i.e., each potential pair  $(\varphi, \varphi^*) \in \tilde{\Theta}$ , the constraint set of  $J(\varphi, \psi)$ .

In fact, we have

*Theorem.* If  $\mu, \nu$  are (Borel) probability measures on  $\mathbb{R}^d$ , with finite second moments, then, with respect to the cost function  $c(x, y) = \frac{1}{2} \|x - y\|^2$ ,

- (i) There exists an potential pair  $(\varphi, \varphi^*)$ , convex conjugate, minimizing  $J(\varphi, \psi)$  on  $\tilde{\Theta}$ ,
- (ii) If, in addition,  $\mu$  is nonatomic, there exists a unique optimal  $\pi^* \in \Pi(\mu, \nu)$  of the form  $\pi^* = \gamma_{T^*} = (I_{\mathcal{X}}, T^*)\#\mu$ , with  $T^* = \nabla\varphi$  unique.

*Comments.*  $T^*$  is the unique minimizer of (MP), with the strict convex loss, which is the gradient of a convex function (hence monotone non decreasing). The optimal potential pair  $(\varphi, \varphi^*)$  is obtained from the dual Kantorovich problem.

Thus, for  $\mu$  being the uniform probability  $du$  on  $\mathcal{X} = [0, 1]^d$ , the (Brenier) map  $T^*$  (pushing  $du$  to  $dF = \nu$ ) is our *multivariate quantile function* of the multivariate distribution function  $F(\cdot)$  on  $\mathbb{R}^d$ .

The convex function  $\varphi$  is not unique, but the gradient  $\nabla\varphi$  is unique ( $\mu - a.e.$ ).

Note also that  $\nu = dF$  could be arbitrary, i.e., having finite second moment or not, in view of McCann’s theorem.

Important: As you have said, we drop the symbol  $\tilde{\varphi}$  for simplicity, the  $\varphi$  in the theorem is really  $\tilde{\varphi}(x) = \frac{\|x\|^2}{2} - \varphi(x)$ , i.e., it is the pair

$$\left( \frac{\|x\|^2}{2} - \varphi(x), \frac{\|y\|^2}{2} - \varphi^*(y) \right)$$

which solves the Monge-Kantorovich problem, and  $\varphi$  is “c-concave” in the sense that it is the function  $x \rightarrow \frac{\|x\|^2}{2} - \varphi(x)$  that is convex. Thus, the explicit formula for  $T^*$  is

$$T^*(x) = x - \nabla\varphi$$

which is the gradient of the convex function  $\frac{\|x\|^2}{2} - \varphi(x)$ . If the cost is  $c(x, y) = h(x - y)$  with  $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  strictly convex, then  $T^*(x) = x - (h')^{-1}(\nabla\varphi(x))$ .

*Some Examples*

(1) Let  $\mu, \nu$  be probability measures of  $\mathbb{R}$ , where  $\mu$  is uniform  $du$  on  $[0, 1]$ , and  $\nu$  is uniform  $d\nu$  on  $[1, 2]$ . Then  $\mu$  is nonatomic, and both  $\mu, \nu$  have finite second moments. Let the cost function be  $c(x, y) = (x - y)^2$ . We are going to verify that the univariate quantile function  $F_\nu^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  of  $F_\nu(\cdot) : \mathbb{R} \rightarrow [0, 1]$  is indeed the unique monotone transport map solving the Monge’s problem (i.e., it is the optimal transport map).



We have  $F_v^{[-1]}(v) = 1 + v$ . It's monotone non decreasing. It is a transport map pushing  $du$  on  $[0, 1]$  to  $dv$  on  $[1, 2]$ . Indeed, let  $F_v^{[-1]}(\cdot) = T(\cdot)$ , and  $a \in [1, 2]$ ,

$$(F_v^{[-1]})\#du((-\infty, a]) = du \circ T^{-1}((-\infty, a]) = du\{u : T(u) \leq a\} = du\{u : 1 + u \leq a\} = du\{u \leq a - 1\} = a - 1 = v(-\infty, a])$$

From theory, we know that such a map  $F_v^{[-1]}(v) = 1 + v$  with the above two characteristic properties is the unique solution of the Monge's problem with respect to the given quadratic loss function, i.e.,

$$T = \arg \min\left\{\int_0^1 (x - S(x))^2 dx : S\#du = dv\right\}$$

so let's verify it. We let  $c(x - y)^2 = h(x - y)$  where  $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ , being  $h(t) = t^2$ , a convex function.

For any transport map  $S(\cdot) : [0, 1] \rightarrow [1, 2]$ , we have

$$M(S) = \int_0^1 (x - S(x))^2 dx = \int_0^1 h(x - S(x)) dx$$

Since  $h(\cdot)$  is convex, we have, by Jensen's inequality ( $h(EX) \leq E(h(X))$ ),

$$M(S) = \int_0^1 h(x - S(x)) dx \geq h\left[\int_0^1 (x - S(x)) dx\right] =$$

$$h\left[\int_0^1 x dx - \int_0^1 S(x) dx\right] = h\left[\int_0^1 x dx - \int_1^2 y dy\right] = h\left(\frac{1}{2} - \frac{3}{2}\right) = h(-1) = 1 =$$

$$\int_0^1 (T(x) - x)^2 dx = M(T)$$

since  $T(x) = x + 1$ . Thus, for any  $S$  such that  $S\#\mu = \nu$ ,  $M(S) \geq M(T)$ .

*Notes.* In the above calculations, since  $S\#du = dv$ , we have  $\int_0^1 S(u) du = \int_1^2 v dv$ .

Thus,  $T(x) = 1 + x = F_v^{[-1]}(x)$  in the above example is optimal for the Monge's problem with quadratic loss:  $M(T) = \min\{M(S) : S\#\mu = \nu\}$ .

The optimal transport map  $T(x) = 1 + x = F_v^{[-1]}(x)$  is unique by Brenier's theorem, since it is a monotone and optimal!. It is the derivative of the convex function  $g(\cdot) : [0, 1] \rightarrow \mathbb{R}$ ,  $g(x) = \frac{1}{2}(1 + x)^2$ .

In general, Brenier's theorem affirms that the unique monotone and optional transport map  $T$ , with respect to the strictly convex  $h(\cdot)$ , is of the form

$$T(x) = x - (h')^{-1}(\nabla\varphi)$$

for an optimal potential pair  $(\varphi, \psi)$  of the Kantorovich dual problem, noting that  $\frac{1}{2}h^2(x) - \varphi(x)$  is the convex function such that  $T = \nabla(\frac{1}{2}h^2(x) - \varphi(x))$ .

In our example, with  $\varphi(x) = -2x$ ,

$$h(x) = x^2 \implies h'(x) = 2x \implies (h')^{-1}(x) = \frac{x}{2} \implies$$

$$T(x) = x - (h')^{-1}(\nabla\varphi) = x - (-2)/2 = x + 1$$

(2) As for optimality in the Kantorovich formulation, here is a simple example.

Let  $\mu = dF, \nu = dG$  be two probability measures on  $\mathbb{R}$ , then the transport map (joint measure with  $\mu, \nu$  as marginals)  $\pi^* = dH$ , where  $H(x, y)$  is the bivariate distribution function  $H(x, y) = H(x) \wedge G(y)$  is the optimal joint measure, i.e., with  $c(x, y) = (x - y)^2$ ,

$$\pi^* = \arg \min \left\{ \int_{\mathbb{R}^2} c(x, y) d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}$$

First, let verify that  $\pi^* = dH$  is indeed in  $\Pi(\mu, \nu)$ . We have

$$dH((-\infty, a] \times (-\infty, b]) = H(a, b) = F(a) \wedge G(b)$$

It follows that

$$dH((-\infty, a] \times \mathbb{R}) = F(a) \wedge G(\infty) = F(a) = dF((-\infty, a])$$

Similarly,  $dH(\mathbb{R} \times (-\infty, b]) = dG((-\infty, b])$ .

Note that, in fact, using copula, it is obvious that  $H(x, y) = H(x) \wedge G(y)$  is a bona fide bivariate distribution function on  $\mathbb{R}^2$ , with marginal distribution functions  $F$  and  $G$ , since  $\mathcal{C}(u, v) = u \wedge v$  is a copula!

Also, in fact, we have  $\pi^* = du \circ (F^{[-1]}, G^{[-1]})^{-1}$ , i.e.,  $\pi^* = (F^{[-1]}, G^{[-1]})\#du$ , where  $du$  is uniform on  $[0, 1]$ . Indeed,

$$du \circ (F^{[-1]}, G^{[-1]})^{-1}((-\infty, a] \times (-\infty, b]) = du\{u : F^{[-1]}(u) \leq a, G^{[-1]}(u) \leq b\} =$$

$$du\{u : u \leq F(a), u \leq G(b)\} = F(a) \wedge G(b) = H(a, b)$$

As such, we have

$$K(\pi^*) = \int_{\mathbb{R} \times \mathbb{R}} c(x, y) d\pi^*(x, y) = \int_0^1 (F^{[-1]}(u) - G^{[-1]}(u))^2 du$$

since, in general, when  $\pi^* = du \circ (F^{[-1]}, G^{[-1]})^{-1}$ , we have for any function  $\zeta(x, y)$ ,

$$\int_{\mathbb{R}^2} \zeta(x, y) d\pi^*(x, y) = \int_0^1 \zeta(F^{[-1]}(u), G^{[-1]}(u)) du = \inf\{K(\pi) : \pi \in \Pi(\mu, \nu)\}$$

The quantity

$$W_2(\mu, \nu) = [\inf\{K(\pi) : \pi \in \Pi(\mu, \nu)\}]^{\frac{1}{2}} = \left[ \int_0^1 (F^{[-1]}(u) - G^{[-1]}(u))^2 du \right]^{\frac{1}{2}}$$

is a *Wasserstein distance* between  $\mu$  and  $\nu$ .

(3) Let  $F$  and  $G$  be two distribution functions on  $\mathbb{R}$ , and let  $H(x, y) = F(x)G(y)$ .

What will be the bivariate quantile function  $Q_H(\cdot, \cdot) : [0, 1]^2 \rightarrow \mathbb{R}^2$  of  $H$ , i.e., monotone and  $Q_H \# du = dH$ , where  $du$  is uniform on  $[0, 1]^2$ .

## 5 Notes on Multivariate Quantile Regression

Like a blessing, one of the inventors of univariate quantile regression, Roger Koenker wrote in his recent paper “Quantile Regression 40 years on” the following about multivariate quantiles:

“...Despite generating an extensive literature, it is fair to say that no general agreement has emerge... in contrast to the sample mean of d-dimensional vectors, there is no consensus about an appropriate notion of multivariate median. In an exciting new development, Carlier, Chernozhukov and Galichon [2] have proposed a vector quantile regression notion motivated by classical Monge-Kantorovich optimal transport theory”.

Armed with the notion of multivariate quantile functions, we elaborate first on its application to regression.

Recall that the notion of unconditional (multivariate) quantile functions is this. Let  $Y$  be a random vector with values in  $\mathbb{R}^d$ . By Lebesgue-Stieltjes’ theorem, the law of  $Y$  is the Borel probability measure  $\nu$  on  $\mathcal{B}(\mathbb{R}^d)$  derived from its distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$  as

$$\nu((-\infty, y]) = dF((-\infty, y]) = F(y)$$

Note that, only when needed that we will call upon the “background” setting: the random vector  $Y$  is “defined” on a probability space  $(\Omega, \mathcal{A}, P)$ , so that  $\nu = PY^{-1}$ , and  $F(y) = P(\omega \in \Omega : Y(\omega) \leq y)$ . In our analysis, the polar factorization  $Y = Q_F(U)$  is more “concrete” to use, where  $U$  is a random vector uniformly distributed on the unit cube  $[0, 1]^d$ , with probability measure denoted as  $du$ , and  $Q_F$  denotes the (multivariate) quantile function.

The quantile function  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$  is the (a.e.) unique monotone (non decreasing) map, such that  $Q_F \# du = dF$ , or equivalently,  $dF(\cdot) = du \circ Q_F^{-1}(\cdot)$ .

In multivariate regression analysis, besides our “target” random vector  $Y$ , we have another random vector  $X$ , taking values in  $\mathbb{R}^k$ , and playing the role of covariates (or regressors) of  $Y$ . As “usual”, we wish to establish a statistical model relating  $Y$  to its covariates  $X$ .

As far as (linear) quantile regression is concerned, the main analysis tool is conditional (multivariate) quantile functions.

It should be noted that the computational aspects in multivariate quantile regression are expected to be much more complicated than the univariate case. Not only the OT framework allowed us to generalize appropriately the univariate case to general case, it provides us with computational methods as well.

As such, to appreciate how OT can help, let’s reformulate univariate conditional quantile analysis in the language of OT.

Let  $Y$  be a real-valued random variable with distribution  $F_Y(\cdot)$ . We keep, in our mind, the abstract setting:  $Y$  is defined on  $(\Omega, \mathcal{A}, P)$ , but focus on its concrete polar factorization  $Y = F_Y^{[-1]}(U)$ .

First, recall that the univariate quantile function  $F_Y^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is the *pseudo-inverse* of the distribution function  $F_Y$  (since, in general,  $F_Y$  is only monotone non decreasing, and right continuous, so that it does not have an inverse) defined as

$$F_Y^{[-1]}(u) = \inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$$

This is well-defined since  $\geq$  is a total order relation on  $\mathbb{R}$ . In fact, the infimum is attained, i.e., the infimum is a minimum.

Some useful properties of  $F_Y^{[-1]}(\cdot)$  are as follows.

a)  $F_Y^{[-1]}(\cdot)$  is *monotone non decreasing*, i.e.,  $u \leq v \implies F_Y^{[-1]}(u) \leq F_Y^{[-1]}(v)$  (Note that, strictly increasing means,  $u < v \implies F_Y^{[-1]}(u) < F_Y^{[-1]}(v)$ ).

*Proof.* For  $u \leq v$ , we have  $\{y \in \mathbb{R} : F_Y(y) \geq v\} \subseteq \{y \in \mathbb{R} : F_Y(y) \geq u\}$  and hence  $\inf\{y \in \mathbb{R} : F_Y(y) \geq v\} \geq \inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$ .

b)  $\inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$  is attained.

*Proof.* “ $\inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$  is attained” means  $F_Y^{[-1]}(u)$  is one of the  $y$  such that  $F_Y(y) \geq u$ , i.e.,  $F_Y(F_Y^{[-1]}(u)) \geq u$ .

For each  $n \geq 1$ , by the definition of infimum, there exists  $y_n \in \mathbb{R}$  such that  $F_Y(y_n) \geq u$  and  $y_n \leq F_Y^{[-1]}(u) + \frac{1}{n}$ .

Since  $F_Y(\cdot)$  is nondecreasing, we then have

$$u \leq F_Y(y_n) \leq F_Y(F_Y^{[-1]}(u) + \frac{1}{n})$$

Next, by right continuity of  $F_Y$ , we have

$$\lim_{n \rightarrow \infty} F_Y(F_Y^{[-1]}(u) + \frac{1}{n}) = F_Y(F_Y^{[-1]}(u))$$

so that  $F_Y(F_Y^{[-1]}(u)) \geq u$ , since for each  $n$ ,  $u \leq F_Y(F_Y^{[-1]}(u) + \frac{1}{n})$ .

c) A *weak representation of  $Y$*  is this. For any random variable  $U$  distributed uniformly on  $[0, 1]$ ,  $Y$  has the same distribution as  $F_Y^{[-1]}(U)$  (written as  $Y \stackrel{D}{=} F_Y^{[-1]}(U)$ ).

*Proof.* It suffices to show that

$$F_Y^{[-1]}(u) \leq y \iff u \leq F_Y(y)$$

since then

$$P(F_Y^{[-1]}(U) \leq y) = P(U \leq F_Y(y)) = F_Y(y)$$

Thus, let’s show the above equivalence.

If  $\omega \in \{\omega \in \Omega : U(\omega) \leq F_Y(y)\}$ , i.e.,  $U(\omega) \leq F_Y(y)$ , then, by definition of  $F_Y^{[-1]}(\cdot)$ ,  $F_Y^{[-1]}(U(\omega)) \leq y$ , and hence

$$\{\omega \in \Omega : U(\omega) \leq F_Y(y)\} \subseteq \{\omega : F_Y^{[-1]}(U(\omega)) \leq y\}$$

Conversely, if  $\omega \in \{\omega : F_Y^{[-1]}(U(\omega)) \leq y\}$ , i.e.,  $F_Y^{[-1]}(U(\omega)) \leq y$ , then  $F_Y(y + \varepsilon) \geq U(\omega)$  for all  $\varepsilon > 0$ , and hence  $F_Y(y) \geq U(\omega)$  by right continuity of  $F_Y$ , so that

$$\{\omega : F_Y^{[-1]}(U(\omega)) \leq y\} \subseteq \{\omega \in \Omega : U(\omega) \leq F_Y(y)\}$$

therefore equality.

**Remark.** (i) By taking set complement, we also have

$$F_Y^{[-1]}(u) > y \iff u > F_Y(y)$$

- (ii) The representation is weak since the equality between  $Y$  and  $F_Y^{[-1]}(U)$  is “in distribution” which is weaker than “almost sure equality”, noting that if  $Y \stackrel{a.s.}{=} F_Y^{[-1]}(U)$  (a “strong representation”, called the polar factorization of  $Y$ ) then  $Y \stackrel{D}{=} F_Y^{[-1]}(U)$ .

d) A *strong representation of  $Y$* . Every time we have a uniform random variable  $V$  on  $[0, 1]$ ,  $Y$  and  $F_Y^{[-1]}(V)$  have the same distribution. If we look at the joint distribution  $\pi_{(Y,V)}$  of  $(Y, V)$ , then we see differences among these variables  $V$  although they all have the same uniform distribution  $du$  on  $[0, 1]$ . Indeed, according to Sklar’s theorem, the joint distribution function  $H(y, v)$  of the random vector  $(Y, V)$  is of the form  $c(F_Y, F_V)$  where  $c$  is a (bivariate) copula. Thus, each  $V$  is in fact determined by its own copula  $c$ , in other words, these  $V$  are indexed by copulas. While they are all in  $\Pi(F_Y, du)$ , the set of all joint distributions with the same marginals  $F_Y, F_U$ , there are different by their associated copulas. Thus, saying that there is a  $U$  with distribution  $du$ , such that  $F_Y^{[-1]}(U) \stackrel{a.s.}{=} Y$ , we mean a special  $V$ , or rather, a special copula  $c^*$  of  $(Y, V)$  such that we actually have  $F_Y^{[-1]}(U) = Y$ .

Let’s elaborate a bit more on “Strong representation”, i.e., equality between random variables in the “almost surely” (with probability one) sense.

The question is: is there a random variable  $V^*$  distributed as  $U$ , i.e., uniformly on  $[0, 1]$  such that  $F_Y^{[-1]}(V^*) \stackrel{a.s.}{=} Y$ ?

The answer is affirmative. Its proof will shed light on how actually to “construct” such a random variable.

*Proof.* Let  $F_Y$  be the distribution function of  $Y$ .

- (i) If  $F_Y$  is strictly increasing, then  $F_Y$  and  $F_Y^{[-1]}$  are bijections with  $F_Y = (F_Y^{[-1]})^{-1}$ . Define  $V^*(\omega) = F_Y(Y(\omega))$ , then  $F_Y^{[-1]}(V^*(\omega)) = Y(\omega)$ , and  $V^*$  is uniform on  $[0, 1]$  since

$$P(\omega : V^*(\omega) \leq u) = P(\omega : F_Y(Y(\omega)) \leq u) =$$

$$P(\omega : Y(\omega) \leq F_Y^{[-1]}(u)) = F_Y(F_Y^{[-1]}(u)) = u$$

(ii) If  $F_Y$  is not strictly increasing, the announced  $V^*$  is constructed as follows.

Let  $A_Y = \{y \in \mathbb{R} : P(\omega : Y(\omega) = y) > 0\} \neq \emptyset$ .

For any  $y \in A_Y$ , define a uniform random variable  $V_y$  on  $\{u \in [0, 1] : F_Y^{[-1]}(u) = y\}$ .

Then define

$$V^*(\omega) = F_Y(Y(\omega))1_{(Y(\omega) \notin A_Y)} + V_{Y(\omega)}1_{(Y(\omega) \in A_Y)}$$

Then  $V^*$  is distributed uniformly on  $[0, 1]$ , and  $F_Y^{[-1]}(V^*) \stackrel{a.s.}{=} Y$ .

Next, you may ask: What does it mean by, say, minimizing an objective function over a collection of random variables? i.e., the solution of the optimization problem is a random variable?

Well, remember how mean linear regression was originated? When predicting a random variable  $Y$  from a covariate  $X$ , using mean squared error, we seek the best random variable built from  $X$ , i.e., minimizing the objective function  $E(Y - \varphi(X))^2$  over all random variables  $Z$  of the form  $\varphi(X)$ , i.e., a function of  $X$ . And, of course, the solution is the special random variable  $E(Y|X)$ .

e) For any distribution function  $F_Y$  on  $\mathbb{R}$ ,  $F_Y \circ F_Y^{[-1]}(u) \geq u$ , for any  $u \in [0, 1]$ ,

If  $F_Y$  is continuous then  $F_Y \circ F_Y^{[-1]}(\cdot) = \text{Identity}$  on  $[0, 1]$ , and  $F_Y(Y)$  is distributed uniformly on  $[0, 1]$ ,

$F_Y$  is continuous if and only if  $F_Y^{[-1]}$  is strictly increasing;  $F_Y$  is strictly increasing if and only if  $F_Y^{[-1]}$  is continuous,

If  $F_Y$  is continuous and strictly increasing then  $F_Y^{[-1]}$  is the inverse of  $F_Y$  :  $(F_Y^{[-1]})^{-1} = F_Y$ .

*A quick recap of univariate quantile regression.*

Let  $X$  be a real-valued random variable with distribution function  $F$ . Unlike moments, quantiles exist for any distributions (heavy-tailed or not). Quantiles are used to define financial risk measures, such as Value-At-Risk which is  $F^{[-1]}(\alpha)$ ,  $(P(X > F^{[-1]}(\alpha)) = 1 - \alpha)$ , and in Linear Quantile regression models.

The  $\alpha$  - quantile  $q_\alpha(F)$  minimizes the objective function

$$a \rightarrow E\rho_\alpha(Y - a) = \int_{\mathbb{R}} \rho_\alpha(y - a)dF(y) = \int_{\mathbb{R}} (y - a)[\alpha - 1_{(-\infty, a)}(y)]dF(y)$$

where

$$\rho_\alpha(u) = u[\alpha - 1_{(u \leq 0)}] = \begin{cases} -(1 - \alpha)u & \text{for } u < 0 \\ \alpha u & \text{for } u \geq 0 \end{cases}$$

i.e.,

$$q_\alpha(F) = \arg \min_a E\rho_\alpha(Y - a)$$

and hence its sample  $\alpha$ - quantile  $q_\alpha(F_n)$  is

$$\arg \min_a \sum_{i=1}^n \rho_\alpha(Y_i - a)$$

leading to the following plausible conditional quantile estimator. Since the conditional  $\alpha$ -quantile  $q_\alpha(Y|X)$  minimizes the LAD loss, i.e., minimizing  $E\rho_\alpha(Y - \varphi(X))$  over all possible  $\varphi(X)$ , if we specify  $q_\alpha(Y|X)$  linearly, i.e.,  $q_\alpha(Y|X) = X\theta(\alpha)$ , then the coefficient  $\theta(\alpha)$  could be estimated by the extremum estimator  $\hat{q}_\alpha(Y|X) = \hat{\theta}(\alpha)$  which is

$$\arg \min_{\theta} \sum_{i=1}^n \rho_\alpha(Y_i - X_i\theta) = \arg \min_{\theta} \sum_{i=1}^n (Y_i - X_i\theta)[\alpha - 1_{(Y_i - X_i\theta < 0)}]$$

for data  $(X_i, Y_i), i = 1, 2, \dots, n$ , drawn from  $(X, Y)$ .

What is important is this. The quantile function  $F^{[-1]}(\cdot)$  (which is left continuous) satisfies the following: For  $a > 0$  and  $b \in \mathbb{R}$ ,

$$q_\alpha(aX + b) = aq_\alpha(X) + b$$

i.e.,

$$F^{[-1]}_{aX+b}(\alpha) = aF^{[-1]}_X(\alpha) + b$$

That is,  $F^{[-1]}(\alpha)$  is *affine equivariant* (the transformation  $x \rightarrow ax + b$  is an affine transformation): the quantile representation of a point after affine transformation agrees with its original quantile representation similarly transformed.

This invariance properly is essential to use the quantile regression

$$Y = \beta_\alpha X + \varepsilon_\alpha$$

since, given  $X$ , suppose we model  $q_\alpha(Y|X) = \beta_\alpha X$ , then we have

$$q_\alpha(Y|X) = q_\alpha(\beta_\alpha X + \varepsilon_\alpha) = \beta_\alpha X + q_\alpha(\varepsilon_\alpha|X) = \beta_\alpha X$$

when we impose the condition  $q_\alpha(\varepsilon_\alpha|X) = 0$ . In other words,

$$Y = \beta_\alpha X + \varepsilon_\alpha \dots \text{with } q_\alpha(\varepsilon_\alpha|X) = 0$$

is equivalent to  $q_\alpha(Y|X) = \beta_\alpha X$ .

Note, however, that unlike the mean, in general,  $q_\alpha(X + Y) \neq q_\alpha(X) + q_\alpha(Y)$

For  $\alpha = \frac{1}{2}$ , the median  $F^{-1}(\frac{1}{2})$  minimizes  $E|X - a|$  over  $a \in \mathbb{R}$ . How about other  $\alpha \in (0, 1)$ ?

**Remark.** We need to figure out an objective function for  $F^{[-1]}(\alpha)$  to minimize also to suggest an *extremum estimator* for it.

Now, observe that the median minimizes also the objective function (risk)  $\frac{1}{2}E|X - a|$  whose loss function is

$$\frac{1}{2}|x - a| = \begin{cases} -\frac{1}{2}(x - a) & \text{for } (x - a) < 0 \\ \frac{1}{2}(x - a) & \text{for } (x - a) \geq 0 \end{cases}$$

or

$$\frac{1}{2}|x - a| = (x - a)\left[\frac{1}{2} - 1_{(x-a < 0)}\right]$$

This observation leads to other loss functions generalizing

$$(x - a)\left[\frac{1}{2} - 1_{(x-a < 0)}\right] = \rho_{\frac{1}{2}}(x - a) = \frac{1}{2}|x - a|$$

where  $\rho_{\frac{1}{2}}(u) = \frac{|u|}{2}$ , by replacing  $\frac{1}{2}$  by an arbitrary  $\alpha \in (0, 1)$  in  $(x - a)\left[\frac{1}{2} - 1_{(x-a < 0)}\right]$ , namely

$$L_{\alpha}(x, a) = \rho_{\alpha}(x - a) = (x - a)\left[\alpha - 1_{(x-a < 0)}\right]$$

Note that  $\rho_{\alpha}(u) = u[\alpha - 1_{(u < 0)}]$  is a nonnegative function.

*Theorem.* The  $\alpha$ -quantile of  $X$  minimizes  $E\rho_{\alpha}(X - a)$  over  $a \in \mathbb{R}$ .

*Proof.* As a function of  $a$ , the objective (associated risk) function

$$\begin{aligned} E\rho_{\alpha}(X - a) &= \alpha[EX - a] - \int_{-\infty}^a (x - a)dF(x) = \\ &= \alpha[EX - a] - \int_{-\infty}^a xdF(x) + a \int_{-\infty}^a dF(x) \end{aligned}$$

is differentiable with (assuming for simplicity that  $F$  is absolutely continuous)

$$\frac{d(E\rho_{\alpha}(X - a))}{da} = -\alpha - a \frac{dF}{dx}(a) + a \frac{dF}{dx}(a) + \int_{-\infty}^a \frac{dF}{dx}(x)dx = F(a) - \alpha$$

Since  $F(\cdot)$  is nondecreasing, the function  $a \rightarrow F(a) - \alpha$  is increasing, so that the function  $a \rightarrow E\rho_{\alpha}(X - a)$  is convex. As such, the first order condition

$$\frac{d(E\rho_{\alpha}(X - a))}{da} = F(a) - \alpha = 0$$

implies that the minimum of  $E\rho_{\alpha}(X - a)$  over  $a$  is attained at  $F(a) = \alpha$ , i.e.,  $a = F^{-1}(\alpha)$ , the  $\alpha$ -quantile of  $F$ . In other words, the  $\alpha$ -quantile  $F^{-1}(\alpha)$  minimizes the risk  $E\rho_{\alpha}(X - a)$  over  $a$ .

**Remark.** Thus, since  $F^{[-1]}(\alpha)$  minimizes  $E\rho_{\alpha}(X - a)$ , its empirical counterpart (sample quantile), namely  $\hat{a}_n = \inf\{x \in \mathbb{R} : F_n^{[-1]}(x) \geq \alpha\}$ , minimizes

$$\int_{\mathbb{R}} \rho_{\alpha}(x - a)dF_n(x) = \frac{1}{n} \sum_{i=1}^n \rho_{\alpha}(X_i - a)$$

Note that, unlike moments, quantiles exist for any kind of distributions including heavy-tailed ones. Note also that, unlike  $(x - a)^2$ , the function  $a \rightarrow \rho_{\alpha}(X - a)$  is not differentiable at any  $a \in \mathbb{R}$ . However, it is continuous and convex.



A similar result for conditional quantiles is this. First, the conditional distribution of  $Y$  given  $X = x$  is  $F_{Y|X=x}(y|x) = P(Y \leq y|X = x) = E[1_{(Y \leq y)}|X = x]$ . Its  $\alpha$ -quantile is simply

$$q_{Y|X}(\alpha) = F_{Y|X}^{[-1]}(\alpha) = \inf\{x \in \mathbb{R} : F_{Y|X}(x) \geq \alpha\}$$

where,

$$F_{Y|X}(y) = P(Y \leq y|X) = E(1_{(Y \leq y)}|X)$$

which always exists, since, for each  $y \in \mathbb{R}$ , the random variable  $1_{(Y \leq y)}$  is bounded, and hence the conditional expectation  $E(1_{(Y \leq y)}|X)$  exists (as a Radon-Nikodym derivative).

*Theorem.* The conditional  $\alpha$ -quantile of  $Y$  given  $X$  minimizes  $E\rho_\alpha(Y - \varphi(X))$  over all possible  $\varphi(X)$ .

*Proof.* Indeed, using the same proof for unconditional quantiles,  $q_\alpha(Y|X = x)$  minimizes  $E[\rho_\alpha(Y - a)|X = x]$  so that (integrating over  $P_X$ ) the function  $x \rightarrow q_\alpha(Y|X = x)$  minimizes  $E\rho_\alpha(Y - \varphi(X))$ . Q.E.D.

For applications, a linear conditional quantile model is

$$Y = \beta(\alpha)X + \varepsilon_\alpha$$

where  $q_{\varepsilon_\alpha|X}(\alpha) = 0$ .

**Remark.** Another useful application of quantiles. The one dimensional notion of quantiles plays an interesting role in connections with *copulas*, *OT*, with applications to *production theory in econometrics*. What is the “rationale” of the *Cobb-Douglas production function*?

Recall that the Cobb-Douglas production function (in one dimensional case) is of the form

$$\Phi(.,.) : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty), \Phi(x, y) = x^a y^b$$

with  $x, y, a, b \geq 0$ .

In econometrics,  $\Phi$  represents the technological relationship between the amount of two inputs such as labor ( $X$ ) and physical capital ( $Y$ ), and the amount of outputs that can be produced by these inputs. In the context of OT, it can model the situation where we wish to assign managers (characterized by scalar characteristic/ talent/  $X$ ) to firms (characterized by their market capitalization  $Y$ ). Consider the case where the number of managers is the same as the number of firms. Of course, an “optimal” assignment is the one which should produce the maximum of outputs (say, surplus).

The economic value generated by a manager with talent  $x$ , when working for a firm with size  $y$ , is the production output  $\Phi(x, y)$ .

Let  $P, Q$  denote the distribution of  $X, Y$ , respectively on  $\mathbb{R}$ . An assignment of managers to firms is a transport map  $T$  such that  $T(X) = Y$ , in distribution. That constraint means that each manager is assigned only to one firm.

The total value created is  $E[\Phi(X, T(X))] = E[XT(X)]$ .

It is intuitive to view an optimal assignment should be such that most talented managers will run largest firms. In other words, the variables  $X, Y$  should be *comonotone* (varying in the same way). This desirable property could be realized when the production function  $\Phi(x, y)$  possesses some appropriate condition.

If we look at the Cobb-Douglas production function, then we see that  $\frac{\partial^2 \Phi(x, y)}{\partial x \partial y} \geq 0$ , a property that we call *supermodularity*.

**Remark.** This property is similar to *affiliation* in the theory of *common value auctions*, where it is reasonable to assume that the bidders' (latent) are called *affiliated*.

This so since it is expected that a high value of one bidder's estimate (of the auctioned object) makes high values of the other estimates more likely.

Now, for a uniform distribution  $U$  on  $[0, 1]$ , we have  $F_P^{[-1]}(U) \sim P$ . This univariate transport is generalized, say, to two dimensions as follows.

Let  $U, V$  be two uniform random variables on  $[0, 1]$ , then for  $\pi \in \mathcal{M}(P, Q)$ , we have

$$(F_P^{[-1]}(U), F_Q^{[-1]}(V)) \sim \pi$$

where the joint distribution of  $(U, V)$  is a copula. And the OT problem is formulated as

$$\sup_{\lambda \in \mathcal{M}(U, V)} E_\lambda[\Phi(F_P^{[-1]}(U), F_Q^{[-1]}(V))]$$

i.e., an extremal copula problem.

Thus,  $X, Y$  are *comonotone* if there is  $U$  uniform on  $[0, 1]$  such that  $X = F_P^{[-1]}(U), Y = F_Q^{[-1]}(U)$ .

It is well known that the copula associated with comonotone variables  $X, Y$  is  $C(u, v) = \min(u, v)$ .

An important theoretical result is this.

**Theorem.** If the surplus (production) function  $\Phi$  is supermodular, then the OT problem

$$\sup_{\pi \in \mathcal{M}(P, Q)} E_\pi[\Phi(X, Y)]$$

has a solution. In particular, if  $P$  has no mass points, then  $F_Q^{-1} \circ F_P(x) = T(x)$  is an optimal transport map satisfied  $Y = T(X)$ .

Looking back at Cobb-Douglas production function, the above result indicates that it is optimal to match higher talented managers to larger firms (and less talented managers to smaller firms).

Just like a complex number  $z = x + iy$  that can be written in polar coordinates as  $z = re^{i\theta}$ , a random variable  $Y$  with distribution  $F$  can be "factored" as  $Y \stackrel{D}{=} F^{-1}(U)$ , called a *polar factorization* of  $Y$ . It is this polar factorization which is the appropriate equivalent representation for univariate quantile function to be extended to higher dimensions, as  $Y = \nabla \varphi(U)$ , when  $Y$  is a random vector in  $\mathbb{R}^d, d \geq 2$ ,  $U$  is uniformly distributed on  $[0, 1]^d$ , and  $\nabla \varphi$  is the gradient of a (unique) convex function  $\varphi : [0, 1]^d \rightarrow \mathbb{R}$ .

Specifically, the vector quantile of a multivariate distribution function  $F$  is the gradient of a convex function, and its justification is within Optimal Transport Theory.

If  $X$  is a  $k$ -dimensional random vector, then the *conditional vector quantile* of  $Y$  given  $X = x$  is the multivariate quantile of the random vector  $Y|X = x$ .

Not only for a parallel with mean linear regression, but in view of natural applications, it seems desirable to extend univariate quantile regression model to *multivariate quantile regression*.

*There are many different approaches to defining the notion of multivariate (vector) quantile, but the BEST one is the (recent, 2016) approach based upon OT that we recommend, and elaborate now.*

Quoting R. Koenker, with respect to multivariate extension of one dimensional quantiles: "...Despite generating an extensive literature, it is fair to say that no general agreement has emerged..." In contrast to the sample mean of  $d$ - dimensional vectors, there is no consensus about an appropriate notion of multivariate median.

**Remark on Orders in  $\mathbb{R}^d$ .** The problem seems to be the lack of a natural total order on  $\mathbb{R}^d$ . The *Pareto* order,  $(x_1, x_2, \dots, x_d) \leq (y_1, y_2, \dots, y_d)$  if and only if  $x_i \leq y_i$  for all  $i = 1, 2, \dots, d$ , is only a partial (but not total) order. The *lexicographic* order (used in dictionary) is a total order on  $\mathbb{R}^d$  where components can be ranked as to importance. It is defined as follows.  $(x_1, x_2, \dots, x_d) \leq (y_1, y_2, \dots, y_d)$  if  $x_1 < y_1$ , or  $x_1 = y_1$  and  $x_2 < y_2$ , or  $x_1 = y_1, x_2 = y_2$  and  $x_3 < y_3$ , or...or  $x_i = y_i, i = 1, 2, \dots, d - 1$  and  $x_d < y_d$ , or  $x_i = y_i, i = 1, 2, \dots, d$ .

Why the problem of extending univariate quantiles to multivariate quantiles so difficult? Well, we have just said it "there is no natural total order relation on  $\mathbb{R}^d$  for  $d > 1$ ".

The extension problem is difficult since we tried to extend the univariate quantile *directly* from its definition. In history of mathematics, often when we face an extension problem, such as fuzzy sets, quantum probability, and even "extension of transport maps to transport plans" in OT (!), while we cannot directly extend an existing notion, we look for some equivalent representation of it which can be extended. In the case of univariate quantile, perhaps mathematicians have this "extension methodology" in mind, but it was not easy to find an equivalent representation of univariate quantile which can be extended.

Finally, the extension problem was found in 2016, thanks to OT! It was R. Koenker himself to announce it.

It is impossible to generalize this one dimensional quantile function to  $\mathbb{R}^d$ , with  $d > 1$ , since there is no (natural) total order relation of  $\mathbb{R}^d$ , if we try to generalize this function so defined. In other words, we cannot "directly" generalize this concept. We could try to generalize it "indirectly"?

Remember, how Kantorovich generalized Monge's OT formulation? For example, how to generalize a permutation  $\sigma$  (a pure assignment) on  $\{1, 2, \dots, n\}$  to transport plan?

We cannot do it "directly", so we search for an equivalent representation of  $\sigma$ , i.e., looking for some indirect way. An equivalent representation (an one-to-one map) of a permutation is a permutation matrix to be generalized.

We could do the same thing to generalize quantiles. Perhaps, the difficulty is to find a "canonical" equivalent representation for the quantile map  $F^{-1}(\cdot)$  which could be extended.

Perhaps, it was so since an equivalent representation of  $F^{[-1]}(\cdot)$  is somewhat “hidden”!

Although we all know that  $F^{[-1]}(\cdot)$  is basic for *simulations* because if  $U$  is a random variable, uniformly distributed on  $[0, 1]$ , then the random variable  $F^{[-1]}(U) = F^{[-1]} \circ U$  has  $F(\cdot)$  as its distribution.

Note again that, while the polar factorization of a random variable is used for simulations, it is somewhat hidden (latent) in quantile regression analysis (not needed).

Thus, a characteristic of  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is that it transports the uniform distribution  $\mathcal{U}$  on  $[0, 1]$  to  $dF$  on  $\mathbb{R}$ , in the “language” of OT, in other words, the quantile function  $F^{[-1]}(\cdot)$  is a transport map in OT theory. Is it an equivalent representation for quantiles? Not obviously!

Any way, what seems to be missing is that the probability space  $([0, 1], \mathcal{U})$  is hidden in the “background”: When we define  $F^{[-1]}(\cdot)$ , we did not (in fact, need) mention it at all. Only its surface after, for simulations.

It is hidden, but it’s there! in the language OT, we need to involve the “background”  $([0, 1], \mathcal{U})$  to describe  $F^{[-1]}(\cdot)$  as a transport map.

So let say this. The quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  is a transport map pushing  $\mathcal{U}$  forward to  $dF$ .

If this is an equivalent representation of  $F^{[-1]}(\cdot)$  in the context of OT, then we hope to be able to say this.

Let  $X : \Omega \rightarrow \mathbb{R}^d$  with multivariate distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ . Then the quantile map of  $F$  is  $Q_F(\cdot) : [0, 1]^d \rightarrow \mathbb{R}^d$ , defined as the transport map pushing forward the uniform probability on  $[0, 1]^d$  to  $dF$  on  $\mathbb{R}^d$ .

We proceed now to justify the above definition of multivariate (vector) quantiles, to specify it, to give meaning to it, to provide examples, to define *conditional multivariate quantiles, and multivariate quantile regression*.

If we look closely at the notion of (univariate) quantile function  $F^{[-1]}$  of a random variable  $X$  with distribution function  $F$ , then we realize something fundamental in Monte Carlo (simulation), namely  $F^{[-1]}(U) \stackrel{D}{=} X$ , for a random variable  $U$ , uniformly distributed on  $[0, 1]$ .

The upshot is this. Rather than “look” at the very definition of  $F^{[-1]}(\cdot)$ , we could “look” at  $F^{[-1]}(\cdot)$  as a map from  $[0, 1]$  to  $\mathbb{R}$ , having the property that  $F^{[-1]}(U) \stackrel{D}{=} X$ .

Specifically, consider  $(\mathcal{X}, \mu) = ([0, 1], u)$ , where  $u$  is the uniform probability measure on  $[0, 1]$ , and  $(\mathcal{Y}, \nu) = (\mathbb{R}, dF)$ . Then we realize that  $F^{[-1]}(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$  is a transport map (of Monge!).

However, in order to say that  $F^{[-1]}(\cdot)$  is characterized by such an OT map, we need to show that it is the only transport map in this OT formulation.

Next, for extending this to the multivariate case, we need to show that in the extended OT formulation, namely  $(\mathcal{X}, \mu) = ([0, 1]^n, u_n)$ , where  $u_n$  is the uniform probability measure on the unit cube  $[0, 1]^n$ , and  $(\mathcal{Y}, \nu) = (\mathbb{R}^n, dF_n)$ , where  $F_n(\cdot)$  is the multivariate distribution function on  $\mathbb{R}^n$ , there is a unique transport map.

If it is so, then *the unique transport map between  $([0, 1]^n, u_n)$  and  $(\mathbb{R}^n, dF_n)$  can be used as the multivariate quantile function of the distribution  $F_n$ .*

It turns out that we do have a theoretical result confirming the above! Thanks to McCann [7].

*Theorem.* (McCann, 1995). Let  $\mu, \nu$  be two probability measures on  $\mathbb{R}^n$ , with  $\mu$  being continuous (i.e., it has no mass points, or equivalently, its associate distribution function is continuous on  $\mathbb{R}^n$ , e.g., uniform measure on unit cube). Then there is a measurable map  $T(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  which is the gradient of some convex function  $\varphi$ , and such that  $\nu = \mu T^{-1}$  (equivalently,  $T(X) = Y$ , where  $X \sim \mu, Y \sim \nu$ ). Moreover,  $T$  is unique  $\mu$ -a.s.

**Remark.** The gradient of a multivariate (differentiable) function ( $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ ) is the vector of its partial derivatives. If  $\varphi$  is differentiable, i.e., having first order partial derivatives, then  $\varphi$  is convex if and only if for any  $x, y \in \mathbb{R}^n$ , we have  $\langle \nabla \varphi(x) - \nabla \varphi(y), x - y \rangle \geq 0$  (gradient monotonicity). For  $n = 1$ , a differentiable convex function has nondecreasing derivative.

The Theorem says that if  $X \sim \mu$ , then there is a unique convex function  $\varphi$  such that its gradient  $\nabla \varphi(X) \sim \nu$ , i.e.,  $\nabla \varphi(\cdot)$  is a transport map.

Let's elaborate a bit on this fundamental theorem.

For  $n = 1$ , consider  $(\mathcal{X}, \mu) = ([0, 1], u)$ , noting that the uniform measure  $u$  is continuous, and  $(\mathcal{Y}, \nu) = (\mathbb{R}, dF)$ . The quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$ , which is non decreasing, and transporting  $u$  to  $dF$ , because  $F^{[-1]}(U) \sim dF$ . The quantile function  $F^{-1}$  is nondecreasing and hence is the derivative of a convex function. Thus,  $F^{[-1]}$  fits perfectly McCann's Theorem, and hence is a (a.s.) unique transport map.

Note that if  $\mu$  is an arbitrary continuous probability measure on  $\mathbb{R}^d$  with associate multivariate distribution function  $F_\mu$ , then the transport map is  $F_\nu^{[-1]} \circ F_\mu(x) = \varphi'(x)$ .

Thus, the univariate quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  with its equivalent representation as a transport map pushing forward the uniform measure on  $[0, 1]$  to the probability measure  $dF$  on  $\mathbb{R}$ , can be extended to higher dimensions, as THE transport map being the gradient  $\nabla \varphi$  of some convex function  $\varphi$  on  $\mathbb{R}^n$  ( $\nabla \varphi$  push forward  $([0, 1]^n, u_n)$  to  $(\mathbb{R}^n, dF_n)$ ).

The  $d$ -quantile function of a multivariate distribution function  $F(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$  is the gradient  $\nabla \varphi : [0, 1]^d \rightarrow \mathbb{R}^d$ , of some convex function  $\varphi : [0, 1]^d \rightarrow \mathbb{R}$ , such that  $\nabla \varphi(U) \sim dF$ , where  $U$  is the uniform random vector on  $[0, 1]^d$ .

The above map  $\nabla \varphi$  (Brenier map) is the map between  $dU$  (uniform probability measure on the unit cube  $[0, 1]^d$ ) and  $dF$ .

In one dimension,  $\nabla \varphi$  is  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  (a nondecreasing function, such that  $F^{-1}(U) \sim F$ ).

Let  $X, Y$  be random vectors on  $\mathbb{R}^d, \mathbb{R}^k$  with distribution  $F, G$ , respectively. Then the conditional multivariate quantile function of  $Y|X = x$  is the Brenier map between  $dU$  on  $[0, 1]^d$  and the conditional probability measure of  $Y|X = x$ , i.e., the multivariate quantile of the conditional distribution.

Specifically, the conditional quantile function of  $Y|X = x$  is  $\nabla \varphi_x$  where  $\varphi_x(\cdot)$  is a convex function on  $[0, 1]^d$  with  $Y = \nabla \varphi_x(U)$ .

Note that there are many attempts to define multivariate quantiles in the literature, but as R. Koenker said, this approach based on OT seems the best! mainly because it capture two basic properties of the univariate quantile function  $F^{[-1]}(\cdot) : [0, 1] \rightarrow \mathbb{R}$  (as a kind of "inverse" of  $F$ , with a precise meaning, e.g., of median), namely  $F^{[-1]}(\cdot)$  is a monotone (nondecreasing) function, and  $F^{[-1]}(U) = Y$  (where  $Y \sim dF$ ). This is

so because, as the gradient of a convex function,  $\nabla\varphi$  is the natural generalization of monotonicity in one dimension case, and  $Y = \nabla\varphi_X(U)$  when  $X$  is a covariate.

Let  $Q_{Y|X}(u|x)$  be the conditional (multivariate) quantile of  $Y|X = x$  at level  $u \in [0, 1]^d$ . A linear model for it is

$$Q_{Y|X}(u|x) = \beta_o(u)^T g(x)$$

so that we have the representation

$$Y = \beta_o(U)^T g(X)$$

with  $U|X \sim \text{uniform } [0, 1]^d$ ,  $\beta(u)$  is  $k \times d$  matrix ( $X \in \mathbb{R}^k$ ).

This formulation leads to a linear programming to computing  $\beta(u)$  both for population and sample settings.

**Remark.** Why do we need to consider multivariate quantile regression?

Well, let's spell it out loud again. At the “beginning”, Gaussian models made statisticians to center their attention only on the mean, and conditional mean of variables of interest. Then it was discovered that linear (univariate) quantile regression can address more issues in economics. However, we only have univariate quantile regression (Koenker & Bassett, 1982). As such, even we are really interested in, say, how household expenditures affect total income, we can only look at a specific component of household expenditures, e.g., food expenditure, one among 9 possible components of household expenditures: Food, Clothing, Housing, Heating and Lighting, Tools, Education, Safety, Medical care, Services. A multivariate ( $d = 9$ ) quantile regression is desirable, and now possible!

Multivariate quantile functions are useful in a variety of fields, see e.g., Galichon [4], Matzkin [6], Panaretos and Zemel [9], Santambrogio [10].

## References

1. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **44**(4), 375–417 (1991)
2. Carlier, G., Chernozhukov, V., Galichon, A.: Vector quantile regression: an optimal transport approach. *Ann. Stat.* **44**, 1165–1192 (2016)
3. Carlier, G., Chernozhukov, V., De Bie, G., Galichon, A.: Vector quantile regression and optimal transport, from theory to numerics. *Empir. Econ.* **62**, 35–62 (2020). <https://doi.org/10.1007/s00181-020-01919-y>
4. Galichon, A.: *Optimal Transport Methods in Economics*. Princeton University Press, Princeton (2016)
5. Koenker, R., Bassett, G.: Regression quantiles. *Econometrica* **46**(1), 33–50 (1978)
6. Matzkin, R.L.: Nonparametric estimation of nonadditive random functions. *Econometrica* **71**(5), 1339–1375 (2003)
7. McCann, R.J.: Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80**(2), 309–323 (1995)
8. Monge, G.: Memoire sur la theorie des dblais et des remblais. *Histoire de l'Academie Royale des Sciences de Paris* 666–704 (1781)

9. Panaretos, V.M., Zemel, Y.: An Invitation to Statistics in Wasserstein Space. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-38438-8>
10. Santambrogio, V.: Optimal Transport for Applied Mathematicians. Birkhauser, Cham (2015)
11. Villani, V.: Topics in Optimal Transportation, vol. 58. American Mathematical Society, Providence (2003)
12. Villani, V.: Optimal Transport: Old and New, vol. 338. Springer, Heidelberg (2008)



# The Unfulfilled Quest for Discovering Cause from Probability

William M. Briggs<sup>(✉)</sup>

13039 Cedar Street, Charlevoix, MI 49720, USA  
matt@wmbriggs.com

**Abstract.** We review the paper “Causal emergence is widespread across measures of causation” by Renzo Comolatti and Erik Hoel, which purports to derive (automatic) mathematical formula, which are not entirely probability based, to discover cause. We disagree with the paper’s central arguments, show where they are flawed, and argue that cause is much more difficult to discover than the authors suspect.

**Keywords:** Cause · Explanation · Probability

## 1 The Quest of Cause

This paper is a review of “Causal emergence is widespread across measures of causation” by Renzo Comolatti and Erik Hoel, [1], an important work in causation in probability and statistics. Important because it neatly and succinctly summarizes the most current mainstream views of how cause might be discovered using statistical and probabilistic methods.

I write for those not convinced by these methods, and present here counter-arguments to Comolatti and Hoel’s claims.

Their paper opens, “While causation has historically been a subject of philosophical debate, work over the last few decades has shown that metaphysical speculations can be put aside in favor of mathematical formalisms.”

I believe this is false: mathematical formalisms cannot with certainty lead to unambiguous cause. It is false at least because you must first have a philosophy of cause, which is to say those metaphysical speculations cannot be dispensed with, and are needed to represent observations mathematically. One can write down any number of mathematical equations, but the act of mapping them onto (what is at least perceived as) Reality, philosophy necessarily enters the discussion. It is a *philosophy* to say equations (i.e. models) can represent Reality, in whatever way they do represent it.

Ignoring what might be construed as bickering over terms, we admit it is logically possible that some philosophy might be mathematically amendable to cause identification. Comolatti and Hoel believe they have found such a formalism.



## 2 The Proposed Formalism

As is usual, they posit  $\Omega$  as the “set of all possible occurrences” of certain events. This is, and must be, conditional on whatever assumptions are made to limit this set (examples below). That is,  $\Omega$  is not and cannot be unconditional, for then it would be the set of all possible events that have ever happened, or ever could happen, which is not a useful set for modeling.

Then comes their philosophy, masked as mathematics: “we can consider causes  $c \in \Omega$  and effects  $e \in \Omega$ , where we assume causes  $c$  to precede effects  $e$ ”. That “precede” is where the difficulty begins, for causes and effects are usually simultaneous. Simultaneous with you turning the page of this paper ( $c$ ), the page on the paper turns ( $e$ ). Both  $c$  and  $e$  happen together.

The authors appear to have in mind time series-like causes and effects only. For example, the grandfather sires the father, who in turn sires the son. The grandfather is in a sense *a* (and not *the*) cause of the grandson. Or consider a string of daily high temperature observations  $y_t$  at some location. Does  $y_{t-1}$  in some way *cause*  $y_t$  because  $y_{t-1}$  precedes it? Classic time series analysis usually only considers this weaker form of cause. See, e.g. [2,3].

Limiting cause to time-series events is, of course, more philosophy; enough that I think it well demonstrated metaphysics are a necessity. It is then, and always, a question of *which* metaphysics.

Now it’s very strange at first to have the causes *inside* the set of all that can happen, though it does appear more natural the “events” live there. Yet this step I, like the authors, believe is necessary, though for different reasons to be discussed later. I rely on the necessary conditional nature of  $\Omega$ , whereas the authors are much vaguer about the constituents of this set. Let us, however, set that aside for the moment and first grasp the essentials of their method.

They let  $c \in C \subseteq \Omega$  for all assumed causes, and  $e \in E \subseteq \Omega$  for all events.  $C$  and  $E$  are thus the set of causes and events, from which individual causes  $c$  and events  $e$  are “found.” It is, or should be, a somewhat disturbing notation, for reasons that will become clear in a moment.

Here is the gist of their philosophy:

As we will see, in order to gauge causation, we will have to evaluate counterfactuals of  $c$ , and consider the probability of obtaining the effect  $e$  given that  $c$  didn’t occur. We will write this probability  $P(e|C \setminus c)$ , where  $C \setminus c$  stand for the complement of  $c$ , by which we mean the probability of  $e$  given that any cause in  $C$  could have produced  $e$  except for  $c$ . Note that although conventionally written  $P(e)$  we will write  $P(e|C)$  to underscore the following notion: namely, that to meaningfully talk about  $P(e|C)$  (and  $P(e|C \setminus c)$ ), a further distribution over  $C$  must be specified. That is:

$$P(e|C) = \sum_{c \in C} P(c)P(e|c)$$

where there is some assumption of a distribution  $P(C)$ .

This is confused. First, there is no such thing, in any context, of an unconditional probability (I prove this in [4]; if you do not know this proof, accept for the moment its soundness for the sake of argument). So we cannot write  $P(c)$  and have it mean

anything. Here begins the trouble. It appears that our authors to represent a cause  $c$  picked from among a set of causes  $C$ .

Some *cause* has to do the picking. Which cause is this? It isn't  $c$  and it isn't in  $C$ , nor  $\Omega$ . It lies outside all assumed knowledge in some mysterious way. Further,  $P(C)$  suffers the same fate of lack of explicit condition, and so adds a second layer of the unknown. From where does  $C$  arise? Who or what *causes* this set of causes? Not only is  $c$  caused to be the cause of  $c$ , in some unidentified way, but something necessarily is causing  $C$ .

Like, say, "randomness". Again like many, they have swept a philosophical assumption under the mathematics and said "No philosophy here! But if something has to pick  $c$ , and something does, we'll call that picking-cause randomness." Which is impossible. Randomness is not a cause. Randomness merely expresses a state of uncertainty about what a cause is.

Second, and a criticism more familiar to a mathematical reader,  $P(e|c)$  must either equal 1 when  $c$  is the cause of  $e$ —for  $c$  has *caused*  $e$  to happen; that is the definition of efficient cause—and  $P(e|c)$  equals 0 when  $c$  is not the cause of  $e$ . The math above still works, though, because if  $C$  is exhaustive of all causes under consideration, then indeed  $P(e|C) = 1$  because just one of  $P(e|c_i) = 1$  and all other  $P(e|c_j) = 0, j \neq i$ .

We can see the trouble. They are confusing *knowledge* of cause with cause itself. There is nothing wrong with, and indeed it is common, of having a supposed set of efficient causes, only one which we assume has operated to cause  $e$ , and all of which are capable of the effect, but we don't know which worked in this particular case, and then using probability to assist in ascertaining the most likely of these causes.

For instance, somebody murdered Mr Body, and it can be one of Colonel Mustard, Professor Plum, Mrs White and so on. The murder is the event, the possible causes  $c$  are the individuals (we're still being loose with the nature of *cause* itself). All six are  $C$ . If we begin with *only*—the word is strict—with the *assumption* it must be one of these six, a *true* assumption conditional on the rules of the game  $g$ , then  $P(c_i|g) = 1/6 \forall i$  and  $P(e|c_i g) = 1 \forall i$ . Thus  $P(e|Cg) = 1$  as required.

In Reality, if an event—a definable observation in the world—happens, there must have been some reason it was so, accepting the principle of sufficient reason (as I do; see [5]). The observation must have been caused. There must be a formal, material, efficient, and "final" cause, i.e. a reason for its existence. This in no way implies we know any or all of these elements in an observation.

Take mysterious lights in the sky. You see them, they happened, but you have no idea why, and don't even care to guess, or aren't experienced enough to guess. But you do notice they seem to happen at somewhat semi-regular intervals. You can *model* the occurrence using probabilities, where the information about intervals goes *inc* (again being loose with *cause*).

It is not as simple as all this, because we cannot consider efficient cause alone. Our authors sort of understand this:

For any cause  $c$ , we can always ask, on one hand, how sufficient  $c$  is for the production of an effect  $e$ . A sufficient relation means that whenever  $c$  occurs,  $e$  also follows...Separably, we can also ask how necessary  $c$  is to bring about  $e$ , that is, whether there are different ways then [*sic*] through  $c$  to produce  $e$ ...Yet these properties are orthogonal: a cause  $c$  may be sufficient to produce  $e$ , and yet there

may be other ways to produce  $e$ . Similarly,  $c$  may only sometimes produce  $e$ , but is the only way to do so.

You can see how confused this is. And this is because the four aspects of cause are not laid out with care. Suppose you have to shoot a man in the head. You do. He dies. You caused the death. But the bullet also caused it—if we're being loose about which part of cause we're discussing, formal, material, efficient, and goal/final. The next man you have to shoot holds up a steel plate which stops the bullet. One form of cause has still operated, but another form has not. The event did not occur, even though two aspects of cause, the material cause of the bullet and the goal of the killing, were present.

So let's call a  $c$  that has all four aspects of cause a *complete cause* of  $e$ . Obviously, then, if  $c$  is the complete cause of  $e$  then  $P(e|c) = 1$ .

They define sufficiency of a cause (not complete) as:

$$\text{suff}(e, c) = P(e|c). \quad (1)$$

And necessity as:

$$\text{nec}(e, c) = 1 - P(e|C \setminus c). \quad (2)$$

To us, if  $c$  is a complete cause, then either  $P(e|c) = 1$  or 0, and nothing else. Given  $c$  is *the* complete cause, then  $1 - P(e|C \setminus c) = 0$  because there are no other complete causes in  $C$ . Or if  $c$  is *a* complete cause among others (another shooter of our victim might also exist and shoot), then again the total is 0, because  $C \setminus c$  still has a complete cause in it.

Nothing has been gained by adding these concepts of sufficiency and necessity. Much is gained in understanding the full or complete cause of any observations, such as you deciding to pull the trigger and the resulting execution.

Again, we must keep clear the difference between cause and *knowledge of cause*. They are not the same. Probability works for the second, but not the first.

Our authors spend time showing how these measures might fit in with other schemes or theories of causality—except the classical Aristotelian view, which we espouse. They create measures of “causal strength” and show how various philosophical theories of cause can be cast as functions of  $\text{suff}(e, c)$  and  $\text{nec}(e, c)$ , and base probabilities like  $P(e|c)$ .

They examine Eells's idea of probability rising, Suppes's similar view, Cheng's idea of causal attribution, Lewis's counterfactual theory, Pearl's various causal measures, Lewis's closest possible worlds, bit-flip measures (examining the change of information measures in a system), and effective information. From all these they compute a “causal strength” (CS) measure, which they show are always functions of  $\text{suff}(e, c)$  and  $\text{nec}(e, c)$ , and of the other probabilities given above, and nothing else.

They begin with Hume's constant conjunction. Briefly, Hume said we only conclude cause because we observe the “constant conjunction” of what we take to be causes and their effects. But Hume used that argument to say that cause could never be known. Hume was *the* arch skeptic, responsible for an endless amount of confusion. See [6, 7]. Our authors seem to think Hume's observation can instead be used to discover cause.

I agree with them on that, not being skeptical of induction as Hume was; though our authors do not appear to grasp that it by induction we learn cause.

At any rate, from Hume's notation of correlation of (their)  $c$  and  $e$ , they derive:

$$CS(e, c) = P(c)P(C \setminus c)[P(e|c) - P(e|C \setminus c)] = P(c)P(C \setminus c)[suff(e, c) + nec(e, c) - 1].$$

They subscript this "Galton", and call it the "Galton measure" of correlation, or rather covariance.

The idea is that this, and all the other causal strength measures the investigate, give positive weight towards more or less likely  $c$ . But we don't need measures like that when we already have probability to give weight to uncertainty beliefs, including about  $c$ . All these other measures attempt, then, to replace probability with decision functions.

Plus, this  $CS$ , and all the others, which are markedly similar to this one, have the same interpretation troubles as the original sufficiency and necessity measures.

I'll again note that it is a fine thing to use probability to pick a most likely cause, or aspect of cause, from an *assumed* set, as the murder mystery example shows. But the assumptions must first be there. It is we who pick the  $C$ . It doesn't matter when, in an inference, the assumptions are made. Probability isn't magic. We can make these assumptions after making the observation, or as the investigation progresses. But we must make them. As we have already made at least *some* assumptions of cause when we make the pertinent observations. It is impossible not to; e.g. [8]. You have to demarcate this thing, this event, that happened somehow. There is no escaping philosophy.

### 3 Counterfactuals

The subject of counterfactuals is important; though not all agree; see [9, 10]. Here's their example:

[Y]ou go away and ask a friend to water your [plant]. They don't, and the plant dies. Counterfactually, if your friend had intervened to water the plant, it'd still be alive, and therefore your friend not watering the plant caused its death. However, if the Queen of England had intervened to water the plant, it'd also still be alive, and therefore it appears your plant's death was caused just as much by the Queen of England. This intuitively seems wrong. How do we appropriately evaluate the space of sensible counterfactuals or states over which we assess causation? As we will discuss, there are several options.

This seeming paradox is caused in the same way many paradoxes are: by forgetting what one has conditioned on; or forgetting that conditioning itself, i.e. the assumptions that must be made, is crucial.

If you begin by assuming the cause under consideration is "friend no water, plant die; friend water, plant live" then the friend not watering, by assumption (the conditions), *caused* the plant to die. Nothing else could have caused the death—even if something else did! This is key. That is, you can have *no knowledge of any other cause*. Because you only allow that one assumption (or condition).

If you instead broaden the conditions to “friend or Queen no water, plant die; friend or Queen water, plant live”. If the plant is dead, *all* we know is that either your friend or the Queen didn’t pass the bucket. You can continue *ad infinitum* here, adding others who might water the plant. It is you who sets the limit. There is no mathematical causal formula which lets you limit this set *a priori*. This is the key.

If, in the original set of assumptions, your friend swears he watered the plant, and you believe him, and the plant died, then you have falsified your assumptions. You may then create new ones. Recall the timing of when assumptions are made, and assignment of probability, does not matter.

So this counterfactual attack, or requirement, including Pearl-like “*do(x)*” operators (see [11]), to understand cause does not work. As always, we must keep in mind the distinction between knowledge of cause and cause itself.

## 4 Emergence

Quite a lot of people like the idea of “emergence”, using analogies of ant colonies and so on (one example from a legion is [12]). Individual ants are tiny-brained and have limited behaviors. Jointly, though, the hive exhibits a high degree of complexity. The hive-level behavior is said to “emerge” from the simple behaviors below, somehow as an entity in and of itself.

This is true in one sense, and useful, but false in the sense that somehow the hive is alive itself, that it *thinks* and behaves and directs the behaviors below it, as an entity itself (apart from trivial causes of the hive itself, like this ant can’t walk where that ant is located). There are no causal powers of the hive that can accomplish this. It is, in the end, only individual ants behaving as the circumstances around them dictate. Explanations of cause (or consciousness) “emerging” from below fail.

I am not certain on what idea our authors are advocating. They have an equation for Causal Emergence (recall *CS* stands for causal strength):

$$CE = CS_{macro} - CS_{micro}$$

If CE is positive, there is causal emergence, i.e., the macroscale provides a better causal account of the system than the microscale. This can be interpreted as the macroscale doing more causal work, being more powerful, strong, or more informative, depending on how the chosen measure of causation is itself interpreted. A negative value indicates *causal reduction*, which is when the microscale gives the superior causal account. Note that the theory is agnostic as to whether emergence or reduction occurs.

If I understand them right, this makes their same mistake of mixing up knowledge of cause and cause itself. This would seem to be a measure of probability usefulness, with very little of actual knowledge of cause about it. And anyway, who decides where the micro stage ends and the macro state begins?

Continue with the ants. A pure probability model at the hive level, for predicting whatever observables about the hive you care to name, might be more useful than semi-causal models of individual ant behavior at predicting hive-level observables. Recall probability can be silent on cause and still be useful. Ask casinos.

But a *complete* causal model of all the ants in the hive must be a superior model to the hive-level probability model. Obviously, if we know everything, we can predict everything. The former model will be huge and complex, and is likely impossible except to gross approximation. The latter probability model can beat the more complex model in cost, time to run, even results.

## 5 Whose Cause

Cause isn't simple. Reaching an understanding of the full cause of an event is unusual for complex events, though quite simple, even trivial, for everyday events.

There are many other automated attempts than the ones examined here. Mentioning just one, [13] believe they have found “deconfounders” using machine learning, which are statistical models with fewer parameters.

I believe all of these methods fail for the same reason, because all discovery of cause come to the same thing. It is we who must specify the set of causes under consideration for any  $e$ , or set of  $e$ , as the *Clue* example above indicates.

Probability can certainly be used to weigh evidence, suggesting one cause is more likely than another, or others. But we still come to knowledge of cause the old-fashioned way: through introspection, intuition, and induction.

## References

1. Comolatti, R., Hoel, E.: Causal emergence is widespread across measures of causation (2022). <https://doi.org/10.48550/ARXIV.2202.01854>
2. Prado, R., West, M.: Time Series: Modeling, Computation, and Inference. Chapman & Hall, London (2010)
3. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer, New York (1991)
4. Briggs, W.M.: Uncertainty: The Soul of Probability, Modeling & Statistics. Springer, New York (2016)
5. Feser, E.: Scholastic Metaphysics: A Contemporary Introduction. Editions Scholasticae, Neunkirchen-Seelscheid, Germany (2014)
6. Stove, D.: Probability and Hume's Inductive Scepticism. Clarendon, Oxford (1973)
7. Stove, D.: The Rationality of Induction. Clarendon, Oxford (1986)
8. Quine, W.V.: Two Dogmas of Empiricism. Harper and Row, Harper Torchbooks, Evanston, IL (1953)
9. Dawid, A.P.: J. Am. Stat. Assoc. **95**(450), 407 (2000). <https://doi.org/10.1080/01621459.2000.10474210>. <https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210>
10. Dawid, A.P.: Stat. Sci. **19**, 44 (2004)
11. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
12. Budenholzer, F.E.: Zygon® **39**(2), 339 (2004). <https://doi.org/10.1111/j.1467-9744.2004.00577.x>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9744.2004.00577.x>
13. Wang, Y., Blei, D.M.: J. Am. Stat. Assoc. **114**(528), 1574 (2019). <https://doi.org/10.1080/01621459.2019.1686987>



# Optimal Transport for Counterfactual Estimation: A Method for Causal Inference

Arthur Charpentier<sup>1</sup>(✉), Emmanuel Flachaire<sup>2</sup>, and Ewen Gallic<sup>2</sup>

<sup>1</sup> Université du Québec à Montréal (UQAM), Montréal (Québec), Canada  
charpentier.arthur@uqam.ca

<sup>2</sup> Aix-Marseille University (Aix-Marseille School of Economics), CNRS, Marseille, France

**Abstract.** Many problems ask a question that can be formulated as a causal question: *what would have happened if...?* For example, *would the person have had surgery if he or she had been Black?* To address this kind of questions, calculating an average treatment effect (ATE) is often uninformative, because one would like to know how much impact a variable (such as the skin color) has on a specific individual, characterized by certain covariates. Trying to calculate a conditional ATE (CATE) seems more appropriate. In causal inference, the propensity score approach assumes that the treatment is influenced by  $\mathbf{x}$ , a collection of covariates. Here, we will have the dual view: doing an intervention, or changing the treatment (even just hypothetically, in a thought experiment, for example by asking what would have happened if a person had been Black) can have an impact on the values of  $\mathbf{x}$ . We will see here that optimal transport allows us to change certain characteristics that are influenced by the variable whose effect we are trying to quantify. We propose here a *mutatis mutandis* version of the CATE, which will be done simply in dimension one by saying that the CATE must be computed relative to a level of probability, associated to the proportion of  $x$  (a single covariate) in the control population, and by looking for the equivalent quantile in the test population. In higher dimension, it will be necessary to go through transport, and an application will be proposed on the impact of some variables on the probability of having an unnatural birth (the fact that the mother smokes, or that the mother is Black).

**Keywords:** Causality · Conditional Average Treatment Effects (CATE) · Counterfactual · Mutatis Mutandis · Optimal transport · Quantiles

## 1 Introduction

### 1.1 From Intervention to Counterfactuals

In Pearl and Mackenzie (2018), a “ladder of causation” is introduced, to describe the three levels of causal reasoning. The first level, named “*association*”, discusses associations (not to use the word “*correlation*”) between variables. Questions such as “*is variable X associated with variable Y?*” can be answered at this level. Econometric models are usually simply based on such associations. The second level is labelled

“intervention”. Reasoning on this level answers questions of the form “if I make the intervention  $T$ , how will this affect the level of the outcome  $Y$ ?” For example, the question “would a patient heal faster at home or at the hospital, after some surgery?” is a standard question on this second level of the ladder of causation. This kind of reasoning invokes causality and can be used to investigate more questions than the reasoning of the first level. The third level of the “ladder of causation” is labelled “counterfactuals” and involves answering questions which ask what might have been, had circumstances been different. An example of counterfactual questions given in the book is “would Kennedy be alive if Oswald had not killed him?”. Counterfactual modeling implies that, to each individual in the control space, described through variables  $\mathbf{x}$  and  $y$ , we will associate a counterfactual version of that individual in the hypothetical space. More formally, we will use notations of causal inference to answer counterfactual questions, such as “would that person have had surgery if she had been Afro-American?”

## 1.2 Causal Inference Framework

Consider, as in Rubin (1974) or Hernán and Robins (2010), the following framework: let  $t$  denote some binary treatment,  $t \in \{0, 1\}$ , with respectively, the control and the treatment. Let  $\mathbf{x}$  be some covariates,  $y$  the observed outcome, with  $y_{T \leftarrow 1}^*$  and  $y_{T \leftarrow 0}^*$  the potential outcomes (also denoted  $y(1)$  and  $y(0)$  in Imbens and Rubin (2015) or Imai (2018), or  $y^1$  and  $y^0$  in Morgan and Winship (2015) or Cunningham (2021), even  $y_{t=1}$  and  $y_{t=0}$  in Pearl and Mackenzie (2018)), realized either under treatment condition ( $t = 1$ ) or under control condition ( $t = 0$ ). Note that the observed outcome is  $y = y_{T \leftarrow t}^*$ , or  $y = t \cdot y_{T \leftarrow 1}^* + (1 - t) \cdot y_{T \leftarrow 0}^*$ . An illustration is reported in Table 1.

We will use the term “treatment” (and letter  $t$ ) even if interventions are not possible, so it is no *per se* a “treatment”. In this article, we try to answer a hypothetical question, like most questions asked at the third level of the “ladder of causality”. For instance, in a context of quantifying discrimination, the “treatment” will denote the sensitive attribute, as in Charpentier (2023), such as the race of an individual, e.g., “what would have been the outcome if that person had been Afro-American?” Since our approach proposes an improvement on the metrics used in causal inference literature, we will use similar notations.

There will be a significant impact of treatment  $t$  on  $y$  if  $y_{T \leftarrow 0}^* \neq y_{T \leftarrow 1}^*$ . More specifically, the causal effect for individual  $i$  is  $\tau_i = y_{i,T \leftarrow 1}^* - y_{i,T \leftarrow 0}^*$ . The average treatment effect (ATE) can be defined as follows:

$$\tau = \text{ATE} = \mathbb{E}[Y_{i,T \leftarrow 1}^* - Y_{i,T \leftarrow 0}^*].$$

Its empirical counterpart, the sample average treatment effect (SATE) writes:

$$\hat{\tau} = \text{SATE} = \frac{1}{n} \sum_{i=1}^n y_{i,T \leftarrow 1}^* - y_{i,T \leftarrow 0}^*.$$

Unfortunately, the later is not directly observable, since one of the two is always missing, but some techniques can be used to provide some robust estimate of that quantity (we will present some of them in the next section).



**Table 1.** Potential outcome framework of causal inference, with one binary treatment  $t_i$ , the observed outcome variable  $y_i$  and the two potential outcomes  $y_{i,T\leftarrow 1}^*$  and  $y_{i,T\leftarrow 0}^*$ , as well as some covariates  $\mathbf{x}_i$ . One of the two potential outcomes is observed, and the other is missing, indicated by the question mark in the table.

	Treatment	Outcome			Age	Gender	Height	Weight
	$t_i$	$y_i$	$y_{i,T\leftarrow 1}^*$	$y_{i,T\leftarrow 0}^*$	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$	$x_{4,i}$
1	1	121	121	?	37	F	160	56
2	0	109	?	109	28	F	156	54
3	1	162	162	?	53	M	190	87

Lastly, in the context of possibly heterogeneous effects, captured through covariates  $\mathbf{x}$  (that can be a subset of the entire set of covariates), the conditional average treatment effect (CATE) is defined as the functional

$$\tau(\mathbf{x}) = \text{CATE}(\mathbf{x}) = \mathbb{E}[Y_{T\leftarrow 1}^* - Y_{T\leftarrow 0}^* | \mathbf{X} = \mathbf{x}]$$

that can be written

$$\tau(\mathbf{x}) = \text{CATE}(\mathbf{x}) = \mathbb{E}[Y_{T\leftarrow 1}^* | \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y_{T\leftarrow 0}^* | \mathbf{X} = \mathbf{x}],$$

as introduced in Hahn (1998) and Heckman et al. (1998). More recently, Hitsch and Misra (2018) used that measure to quantify heterogeneous treatment effects to evaluate optimal targeting policies, as well as Powers et al. (2018) and Fan et al. (2022). Wager and Athey (2018), Athey and Wager (2019) and Athey et al. (2019) suggested to use random forests to estimate this quantity, inspired by Davis and Heller (2017). See also Künzel et al. (2019) or Hsu et al. (2022) for additional discussion on that quantity.

A classical assumption is that  $(t_i, y_i, \mathbf{x}_i)$  is a random sample of size  $n$  from some joint random vector  $(T, Y, \mathbf{X})$ . Rosenbaum and Rubin (1983) suggested a strong “ignorable treatment assignment” assumption defined as a conditional independence between  $(Y_{T\leftarrow 0}^*, Y_{T\leftarrow 1}^*)$  and  $T$ , conditional on the covariates  $\mathbf{X}$ .

### 1.3 Agenda

In Sect. 2, and more specifically in Sect. 2.1, we will discuss further the (possible) connection between covariates  $\mathbf{x}$ , treatment  $t$  and the outcome  $y$ . Following our example on discrimination, the treatment variable  $t$  (such as the skin color) is an “exogenous variable”, in the sense that it cannot be influenced either by covariates  $\mathbf{x}$  or by the outcome  $y$ . Using the terminology from directed acyclic graphs (DAGs),  $t$  will have no parent, so in a sense, it will be easier to pretend that an hypothetical intervention on  $t$  is possible. In most applications,  $t$  will have an impact on the outcome  $y$ , but not only. More precisely, it is possible that  $t$  might influence some covariates  $\mathbf{x}$ , and those covariates can, in turn, impact the outcome  $y$ . In Sect. 2.2, we suggest an extension from the standard *ceteris paribus*  $\text{CATE}(\mathbf{x})$  defined as the difference  $\mathbb{E}[Y_{T\leftarrow 1}^* | \mathbf{x}] - \mathbb{E}[Y_{T\leftarrow 0}^* | \mathbf{x}]$ , to some *mutatis mutandis*  $\text{CATE}(\mathbf{x})$  defined as the difference  $\mathbb{E}[Y_{T\leftarrow 1}^* | \mathbf{x}_{T\leftarrow 1}] - \mathbb{E}[Y_{T\leftarrow 0}^* | \mathbf{x}]$ , where, if  $\mathbf{x}$

is considered with respect to the control group, the counterfactual in the treated population should be based on a different version of  $\mathbf{x}$ , in the treated space. As discussed in Sect. 2.3, the classical tool used in econometrics is the propensity score, based on  $\mathbb{P}[T = 1 | \mathbf{X} = \mathbf{x}]$ , that is usually considered to take into account the association that exists between the treatment and the covariates. At the second stage of the “ladder of causation” –the intervention– we consider the fact that  $\mathbf{x}$  might influence  $t$ . When answering the question “*would a patient heal faster at home or at the hospital, after some surgery?*”, it might be relevant to assume that the propensity score can be used to correct for the bias we have in the data, since some patient have been healing at the hospital, not by choice, but because of some  $\mathbf{x}$ . At the third stage of the ladder –the counterfactuals– some sort of dual version should be considered, since  $t$  is not influenced by  $\mathbf{x}$ , quite the opposite: some  $\mathbf{x}$  might be influenced by  $t$ . A simple toy example, based on a Gaussian structural equation model (SEM), is presented in Sect. 2.4, while in Sect. 2.5, we briefly present real data that we will use in the next sections to illustrate various algorithms, based on births in the United States. The variable of interest  $y$  is a binary variable, indicating whether a birth was natural, or not. The covariates  $\mathbf{x}$  considered here will be the weight of the newborn, and the weight gain of the mother. And various “treatments” are considered: whether the mother is Afro-American, or not; whether the mother is a smoker, or not; whether the baby is a girl, or not (results for the last two are reported in Appendix 5.2).

In Sect. 3, we will focus on the case where only one covariate  $x$  is considered. We will start with classical matching techniques in Sect. 3.1, used to match each point in  $(y_i, x_i, t_i = 0)$  –in the control group– with another one in  $(y_j, x_j, t_j = 1)$  –in the treated group– when the two groups have the same size. In Sect. 3.2, we will suggest an “optimal” matching algorithm, to associate individual  $i$  (in the control group) to  $j$  (in the treated group), that we will denote  $j_i^*$ . Then, in Sect. 3.3, we will discuss the case where the two groups have different sizes, that will be called optimal “coupling”. In Sect. 3.4, we will define an estimator, the *mutatis mutandis* CATE,  $\widehat{m}_1(\widehat{\mathcal{F}}(x)) - \widehat{m}_0(x)$ , where  $\widehat{\mathcal{F}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ , with  $\widehat{F}_0$  and  $\widehat{F}_1$  denoting the empirical distribution functions of  $x$  conditional on  $t = 0$  and  $t = 1$ , respectively. We will use quantiles to optimally “transport”  $x$ ’s from the control group to the treated group, formally through the  $\mathcal{T}$  mapping. Finally, in Sect. 3.5, we will illustrate this on probability to have a non-natural baby delivery, on our dataset.

In Sect. 4, we will extend our previous approach to the case where several covariates  $\mathbf{x}$  are considered. Formally, we will use optimal transport techniques to get a proper counterfactual of  $\mathbf{x}$ , not in the control group, but in the treated group. In Sect. 4.1, we will define the optimal transport problem for any number of dimensions and then, in Sect. 4.2, we will explain how to optimally associate each observation  $\mathbf{x}_i$  in the control group (when  $t = 0$ ) with a single counterfactual observation  $\mathbf{x}_j$  in the treated group (when  $t = 1$ ) when the two groups have the same size. This can be related to the Gaussian SEM discussed in Sect. 2.4. In Sect. 4.3, we will see the extension when the two groups have different sizes. Unfortunately, those approach does not provide an explicit mapping  $\mathcal{T}$ , but simply a matching of a single individual  $\mathbf{x}_i$  (in the control group) to a weighted sum of multiple  $\mathbf{x}_j$  (in the treated group). As we will see in Sect. 4.4, it will be possible to get explicit formulation for the mapping  $\mathcal{T}$  (from the space of covariates

in the control group to the space of covariates in the treated group) when we assume that  $\mathbf{X}$  conditional on  $T$  has Gaussian distributions. In Sect. 4.5, those techniques will be further discussed in the context of the application to non-natural birth.

## 2 Ceteris Paribus vs. Mutatis Mutandis

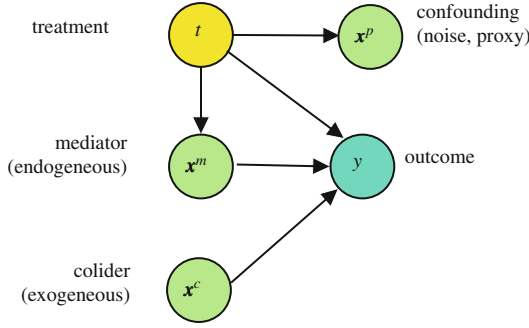
Before introducing another concept of CATE, we will formalize a little bit more the connections between the “treatment”  $t$ , the outcome  $y$  and the covariates  $\mathbf{x}$ .

### 2.1 Exogeneity, Endogeneity and Causal Graphs

As discussed earlier, when presenting the second stage of the “ladder of causation”,  $t$  is a treatment. For example, in epidemiology,  $t$  may be a treatment given to patients, possibly resulting from an intervention. At the third level, the treatment would be more a thought experiment (the “*gedankenexperiment*” in Mach (1893)), to answer a question such as “*what if  $t$  had taken another value?*”, without being able to make an experiment. Chisholm (1946) introduced the idea of “*contrary-to-fact conditional*”, coined as “*counterfactual*” in Goodman (1947). A classical example would be when  $t \in \{\text{smoker, non-smoker}\}$ , since it is not ethically possible to force someone to smoke, but it can also be used on inherent variables, such as the gender or the race of a person, that cannot be changed in a real experiment, to quantify possible discrimination.

Covariates  $\mathbf{x}$  are available variables that have an impact on the outcome  $y$ . It is necessary here to distinguish two kinds of covariates, with variables that are influenced by the value of  $t$ , that might be seen as “endogenous”, and those that are not influenced by the value of  $t$ , that might be seen as “exogenous”. For example, the weight of the baby  $x$  is an endogenous variable with respect to the variable indicating whether the mother is a smoker or not. Using a terminology used on causal graphs, “endogeneous” covariates  $x$  are mediator variables (between  $t$  and  $y$ ), while “exogeneous” ones are variables colliding with  $t$  on  $y$ , sometimes called collider variables (see Fig. 1).

The Markov assumption, on causal networks, states that each variable is conditionally independent of its non-descendants, given its parents. In Fig. 1, in the ‘cofounder’ case (with the fork  $t \rightarrow x$  and  $t \rightarrow y$ ), and in the ‘mediator’ case (with the chain  $t \rightarrow x \rightarrow y$ ),  $y$  is independent of  $t$ , conditional on  $x$ . But in the “Collider” case (with  $x \rightarrow y$  and  $t \rightarrow y$ ), while  $x$  and  $t$  are independent, they become conditionally dependent, conditional on  $y$ . We will not discuss here the construction of the causal graphs, that is supposed to be given (see, e.g., Vowels et al. (2021) for a survey on techniques used to discover causal structures).



**Fig. 1.** Distinction of covariates, with confounding variables that will not influence  $y$  on the top right, and two sets of explanatory variables that will influence  $y$ , that are influenced, or not, by “treatment”  $t$ , with mediators and colliders, at the bottom left.

### 2.2 Impact of a Treatment $t$ on $y$ and $x$ , and CATE

Consider some treatment  $t$ . Let  $x^m$  denote the set of mediator variables and  $x^c$  denote the set of collider variables, as in Fig. 2. Following the SEM terminology used in causal inference, consider data generated according to the equations on the left below (real world), prior to intervention on  $t$ . The right hand equations describe the data generating process with an intervention on  $t$  (denoted  $do(t)$  in Pearl and Mackenzie (2018)):

$$\begin{array}{cc}
 \text{real world} & \text{with intervention } (do(t)) \\
 \left\{ \begin{array}{l} T = h_t(U_t) \\ \mathbf{X}^m = h_m(T, \mathbf{U}_m) \\ \mathbf{X}^c = h_c(\mathbf{U}_c) \\ Y = h_y(T, \mathbf{X}^m, \mathbf{X}^c, U_y) \end{array} \right. & \left\{ \begin{array}{l} T = t \\ \mathbf{X}_{T \leftarrow t}^m = h_m(t, \mathbf{U}_m) \\ \mathbf{X}^c = h_c(\mathbf{U}_c) \\ Y_{T \leftarrow t} = h_y(t, \mathbf{X}_{T \leftarrow t}^m, \mathbf{X}^c, U_y) \end{array} \right.
 \end{array}$$

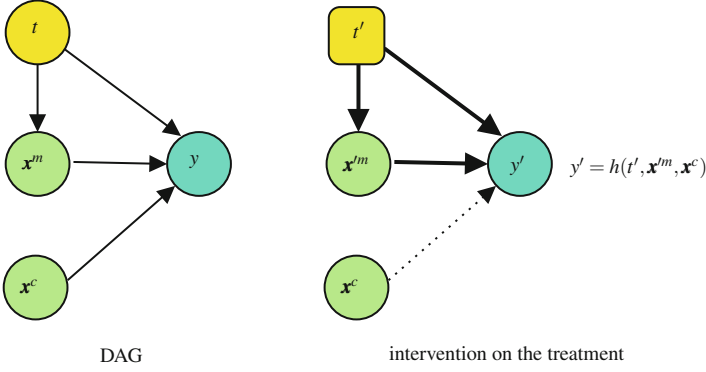
Consider some independent noise variables  $\{U_t, \mathbf{U}_m, \mathbf{U}_c, U_y\}$  (that can be assumed to be centered Gaussian to be close to the econometric literature). In the “real world”,  $T$  is a function of  $U_t$ , and  $U_t$  only, through some  $h_t : \mathbb{R} \rightarrow \{0, 1\}$  function,  $h_t(u) = \mathbf{1}(u > \text{threshold})$ . Then we have two possible explanatory variables: mediator (endogenous) and collider (exogenous). If  $\mathbf{X}^c$  are functions of the noise  $\mathbf{U}_c$  only (through function  $h_c$ ),  $\mathbf{X}^m$  are functions of the noise  $\mathbf{U}_m$  and the treatment  $T$  (through function  $h_m$ ). And finally, the outcome  $Y$  is function of  $\mathbf{X}^c$  and  $\mathbf{X}^m$ , also possibly  $T$ , and some idiosyncratic noise  $U_y$ .

In a *ceteris paribus* approach, CATE( $x$ ) is equal to  $\mathbb{E}[Y_{T \leftarrow 1}^* | x] - \mathbb{E}[Y_{T \leftarrow 0}^* | x]$ . In a *mutatis mutandis* version, we should not consider  $x$ , but a version of  $x$  that should be influenced by the treatment  $t$ , denoted  $x_{T \leftarrow 1}$ . In a general setting, we have the following definition:

**Definition 1.** The *mutatis mutandis* CATE is

$$\text{CATE}(\mathbf{x}) = \mathbb{E}[Y_{T \leftarrow 1}^* | \mathbf{x}_{T \leftarrow 1}] - \mathbb{E}[Y_{T \leftarrow 0}^* | \mathbf{x}]$$

(we might denote  $\mathbf{x}_{T \leftarrow 0}$  instead of  $x$  to avoid confusion for the second term).



**Fig. 2.** A causal graph on the left, and the impact of an intervention on the treatment  $t$  on the right.

More specifically, when we ask the question “*what would have been the probability to have a non-natural delivery for a baby with weight  $x$  if the mother had been smoking?*”, we have to take into account the fact that if the mother had been smoking, the weight of the baby would have been impacted. The original weight  $x$ , associated with a non-Black mother, would become  $\mathbf{x}_{T \leftarrow 1}$  (instead of  $x$ ) if we seek a counterfactual version of  $x$  in the treated population.

### 2.3 Propensity Score Weighting

The classical approach in causal inference is based on the idea that  $T$  is not really exogenous, and can be influenced by  $\mathbf{x}$ . Therefore, the average treatment effect ATE =  $\mathbb{E}[Y_{T \leftarrow 1}^* - Y_{T \leftarrow 0}^*]$ , that can be written

$$\text{ATE} = \mathbb{E} \left[ \frac{TY}{p(\mathbf{X})} - \frac{(1-T)Y}{1-p(\mathbf{X})} \right]$$

would be estimated by

$$\text{SATE} = \frac{1}{n} \sum_{i=1}^n \frac{t_i y_i}{\widehat{p}(\mathbf{x}_i)} - \frac{(1-t_i) y_i}{1-\widehat{p}(\mathbf{x}_i)},$$

where  $p(\mathbf{x})$  is a “propensity score” defined as  $p(\mathbf{x}) = \mathbb{P}[T = 1 | \mathbf{X} = \mathbf{x}]$ , that can be estimated using, for instance, a logistic regression

$$\widehat{p}(\mathbf{x}) = \frac{\exp[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}]}{1 + \exp[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}]}.$$

Thus, the SATE can be seen as the difference between two weighted averages of  $y_i$ 's. As discussed in Abrevaya et al. (2015), it can be used to estimate CATE( $x$ ), on a subset of features, with a local estimate of the average

$$\text{CATE}(x) = \frac{1}{\sum K_h(x_i - x)} \sum \left( \frac{t_i y_i}{\widehat{p}(\mathbf{x}_i)} - \frac{(1-t_i) y_i}{1-\widehat{p}(\mathbf{x}_i)} \right) K_h(x_i - x),$$

using some kernel function  $K_h$ . A  $k$ -nearest neighbors estimate can also be considered:

$$\text{CATE}(x) = \frac{1}{k} \sum_{i \in \mathcal{Y}_k(x)} \left( \frac{t_i y_i}{\widehat{p}(\mathbf{x}_i)} - \frac{(1-t_i)y_i}{1-\widehat{p}(\mathbf{x}_i)} \right),$$

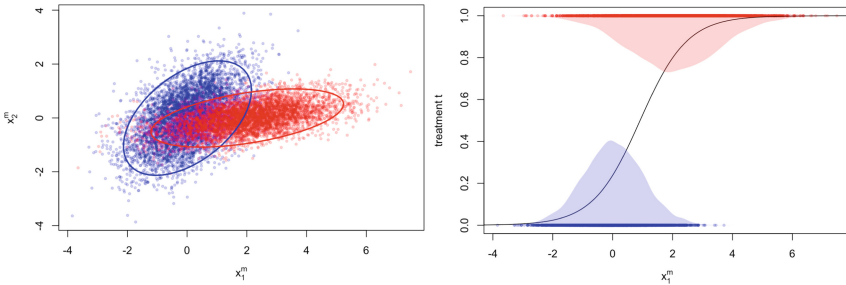
where  $i \in \mathcal{Y}_k(x)$  when  $x_i$  is among the  $k$ -nearest neighbors of  $x$ . If  $y$  is binary (as the example we will use later on), the ATE is a difference between two probabilities, and logistic regressions can be used to properly estimate  $\mathbb{E}[Y_{T=t}^* | \mathbf{X} = \mathbf{x}]$ , with weights in the regressions, that would be either the inverse of  $1 - \widehat{p}(\mathbf{x}_i)$  if  $t_i = 0$  or the inverse of  $\widehat{p}(\mathbf{x}_i)$  if  $t_i = 1$ , as in Li et al. (2018).

### 2.4 A Toy (Gaussian) Example

To illustrate our approach, as an alternative to the use of a propensity score, consider the following toy example, with three explanatory variables, two endogenous (and correlated) ones, and an exogenous one, with some linear model (a Gaussian structural equation model, SEM):

$$\begin{cases} T = \mathbf{1}(U_t < 0), U_t \sim \mathcal{N}(0, 1) \\ \mathbf{X}^m = \boldsymbol{\mu}_T + \boldsymbol{\Sigma}_T^{1/2} \mathbf{U}_m, \mathbf{U}_m \sim \mathcal{N}(\mathbf{0}, \mathbb{I}) \\ X^c = \mu + \sigma U_c, U_c \sim \mathcal{N}(0, 1) \\ Y = \alpha + (\boldsymbol{\beta}_m, \beta_c)(\mathbf{X}^m, X^c)^\top + \gamma T + U_y, U_y \sim \mathcal{N}(0, 1) \end{cases} \quad (1)$$

where all the noises ( $U_t, \mathbf{U}_m, U_c, U_y$ ) are assumed to be centered, and independent. Here  $\boldsymbol{\Sigma}_0^{1/2}$  is Cholesky decomposition of  $\boldsymbol{\Sigma}_0$ , so that  $\mathbf{X}^m$  conditional on  $T = t$  has distribution  $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ . Treatment  $T$  is a binary variable, well-balanced since  $\mathbb{P}(T = 0) = \mathbb{P}(T = 1)$ . Conditional on  $T = t$ , the mediator (endogeneous) variables  $\mathbf{X}^m$  have a Gaussian distribution, with mean  $\boldsymbol{\mu}_t$  and variance matrix  $\boldsymbol{\Sigma}_t$ . A collider variable  $X^c$  is supposed to be independent of the other ones. And finally,  $Y$  is a Gaussian variable where the average is a linear combination of  $\mathbf{X}^m$  and  $X^c$ , plus  $\gamma$  when  $T = 1$ . In Fig. 3, the left-hand panel shows a scatter plot of  $\mathbf{x}^m = (x_1^m, x_2^m)$  with blue points when  $t = 0$  and red points when  $t = 1$ . The right-hand panel shows  $(x_1^m, t)$  on a scatter plot, with the two conditional densities, as well as the logistic regression of  $t$  against  $x_1^m$  (that could be seen as the propensity score).



**Fig. 3.** Scatter plot of  $\mathbf{x}^m = (x_1^m, x_2^m)$  with blue points when  $t = 0$ , and red points when  $t = 1$ , on the left, and the logistic regression of  $t$  against  $x_1^m$  on the right. Toy dataset generated from Eq. (1).

The two interventions yield

$$\begin{array}{l} do(T=0) \\ \left\{ \begin{array}{l} T \leftarrow 0 \\ \mathbf{X}^m = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_0^{1/2} \mathbf{U}_m \\ X^c = \mu + \sigma U_c \\ Y = \alpha + (\boldsymbol{\beta}_m, \beta_c)(\mathbf{X}^m, X^c)^\top + U_y \end{array} \right. \end{array} \quad \begin{array}{l} do(T=1) \\ \left\{ \begin{array}{l} T \leftarrow 1 \\ \mathbf{X}^m = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_1^{1/2} \mathbf{U}'_m \\ X^c = \mu + \sigma U'_c \\ Y = \alpha + (\boldsymbol{\beta}_m, \beta_c)(\mathbf{X}^m, X^c)^\top + \gamma + U'_y \end{array} \right. \end{array}$$

more precisely, in that model with three covariates,  $\mathbf{X}^m = (X_1^m, X_2^m)$ , and since

$$\boldsymbol{\Sigma}_t = \begin{pmatrix} \sigma_{t1}^2 & r_t \sigma_{t1} \sigma_{t2} \\ r_t \sigma_{t1} \sigma_{t2} & \sigma_{t2}^2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_t^{1/2} = \begin{pmatrix} \sigma_{t1} & 0 \\ \sigma_{t2} r_t & \sigma_{t2} \sqrt{1 - r_t^2} \end{pmatrix}$$

we can write

$$\begin{array}{l} do(T=0) \\ \left\{ \begin{array}{l} T \leftarrow 0 \\ X_1^m = \mu_{01} + \sigma_{01} U_1^m \\ X_2^m = \mu_{02} + \sigma_{02}(r_0 U_1^m + \sqrt{1-r_0^2} U_2^m) \\ X^c = \mu + \sigma U_c \\ Y = \alpha + \beta_1^m X_1^m + \beta_2^m X_2^m + \beta^c X^c + U_y \end{array} \right. \end{array} \quad \begin{array}{l} do(T=1) \\ \left\{ \begin{array}{l} T \leftarrow 1 \\ X_1^m = \mu_{11} + \sigma_{11} U_1^{m'} \\ X_2^m = \mu_{12} + \sigma_{12}(r_1 U_1^{m'} + \sqrt{1-r_1^2} U_2^{m'}) \\ X^c = \mu + \sigma U'_c \\ Y = \alpha + \beta_1^m X_1^m + \beta_2^m X_2^m + \beta^c X^c + \gamma + U'_y \end{array} \right. \end{array}$$

and therefore

$$\begin{cases} Y_{T \leftarrow 0} = \alpha + \beta_1^m x_1 + \beta_2^m (\mu_{02} + \sigma_{02} r_0 \sigma_{01}^{-1} [x_1 - \mu_{01}] + \sqrt{1-r_0^2} U_2^m) + \beta^c (\mu + \sigma U_c) + U_y \\ Y_{T \leftarrow 1} = \alpha + \beta_1^m x'_1 + \beta_2^m (\mu_{12} + \sigma_{12} r_1 \sigma_{11}^{-1} [x'_1 - \mu_{11}] + \sqrt{1-r_1^2} U_2^{m'}) + \beta^c (\mu + \sigma U'_c) + \gamma + U'_y \end{cases}$$

Hence,

$$\text{ATE} = \mathbb{E}[Y_{T \leftarrow 1} - Y_{T \leftarrow 0}] = \gamma.$$

For conditional average treatment effects,

$$\begin{cases} \mathbb{E}[Y_{T \leftarrow 0} | X_1^m = x_1] = \alpha + \beta_1^m x_1 + \beta_2^m (\mu_{02} + \sigma_{02} r_0 \sigma_{01}^{-1} [x_1 - \mu_{01}]) + \beta^c \mu \\ \mathbb{E}[Y_{T \leftarrow 1} | X_1^m = x'_1] = \alpha + \beta_1^m x'_1 + \beta_2^m (\mu_{12} + \sigma_{12} r_1 \sigma_{11}^{-1} [x'_1 - \mu_{11}]) + \beta^c \mu + \gamma \end{cases}$$

*Ceteris paribus*, we suppose that  $x'_1 = x_1$ , then

$$\text{CATE}_{cp}(x_1) = \mathbb{E}[Y_{T \leftarrow 1} | X_1^m = x_1] - \mathbb{E}[Y_{T \leftarrow 0} | X_1^m = x_1] = \text{ATE} + \delta x_1 + \kappa,$$

where

$$\begin{cases} \kappa = \beta_2^m (\mu_{12} + \sigma_{02} r_0 \sigma_{01}^{-1} \mu_{01} - \sigma_{12} r_1 \sigma_{11}^{-1} \mu_{11} - \mu_{02}) \\ \delta = \beta_2^m (\sigma_{12} r_1 \sigma_{11}^{-1} - \sigma_{02} r_0 \sigma_{01}^{-1}) \end{cases}$$

*Mutatis mutandis*, since  $X_1^m = \mu_{01} + \sigma_{01} U_1^m$  when  $T = 0$  while  $X_1^m = \mu_{11} + \sigma_{11} U_1^{m'}$  when  $t = 1$ , it is legitimate to consider that  $x'_1 = x_{1:T \leftarrow 1} = \mu_{11} + \sigma_{11} (\sigma_{01}^{-1} [x_1 - \mu_{01}])$ , and therefore, *mutatis mutandis*,

$$\text{CATE}_{mm}(x_1) = \text{ATE} + \delta' x_1 + \kappa',$$

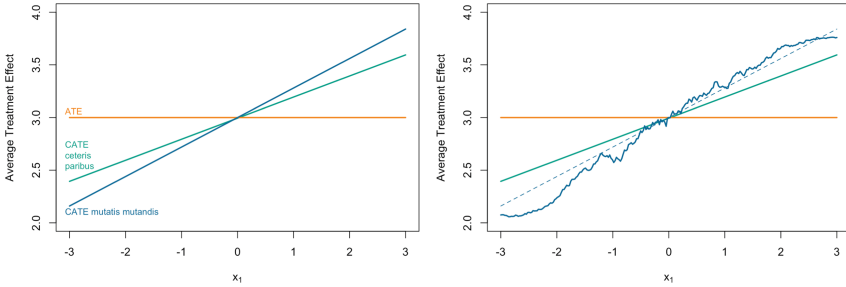
where

$$\begin{cases} \kappa' = \kappa + \beta_1^m [\mu_{11} - \sigma_{11} \sigma_{01}^{-1} \mu_{01}] \kappa + k \\ \delta' = \delta + \beta_1^m (\sigma_{11} \sigma_{01}^{-1} - 1) = \delta + d \end{cases}$$

so that we can also write

$$\text{CATE}_{mm}(x_1) = \text{CATE}_{cp}(x_1) + (dx_1 + k).$$

In Fig. 4, the horizontal orange line is the true average treatment effect (ATE). The green line is the true *ceteris paribus* CATE, while the blue line is the true *mutatis mutandis* CATE, both function of  $x_1^n$ . The dashed and erratic lines on the right-hand graph are estimations of the CATE function using two techniques, described in the next section.



**Fig. 4.** ATE, *ceteris paribus*  $\text{CATE}_{cp}(x_1)$  and *mutatis mutandis*  $\text{CATE}_{mm}(x_1)$  on the left, with an estimate of *mutatis mutandis*  $\text{CATE}_{mm}(x_1)$  on the right, from the toy dataset from example 1. Numerical details are given in Appendix 5.1.

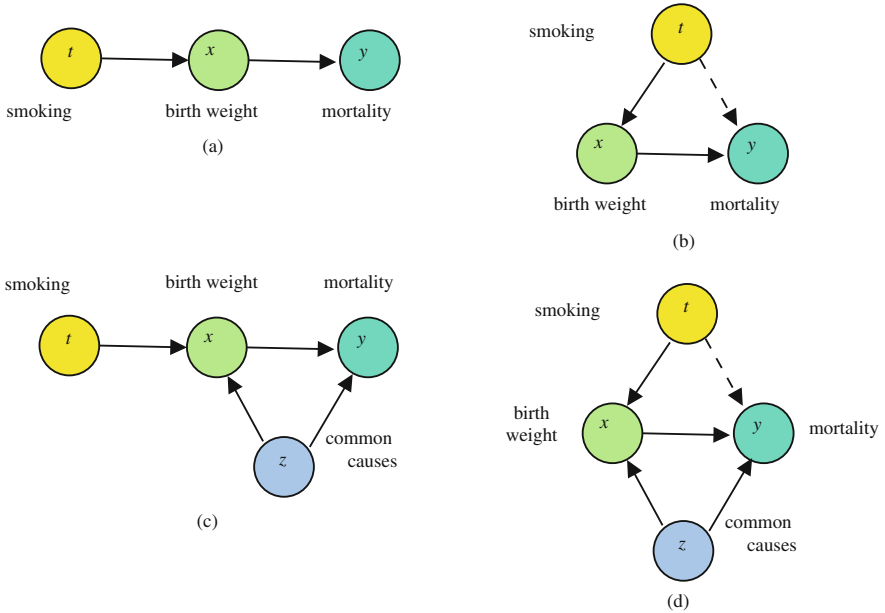
### 2.5 Application on Birth Data

Let us now consider the dataset of all deliveries in the U.S. in 2013.<sup>1</sup> Those data have been intensively used to discuss the “low birth weight paradox”. As explained in Wilcox (1993; 2001), low birth weight of babies  $x$  is strongly associated with increased neonatal mortality  $y$ . However, low birth weight infants born to mothers who smoke  $t = 1$  usually have lower mortality rates than low birth weight infants born to nonsmoking mothers  $t = 0$ . Hernández-Díaz et al. (2006) discussed the birth weight paradox based on causal directed acyclic graphs as a conceptual framework. Multiple causal models have been considered. Figure 5 illustrates four situations, using directed acyclic graphs. In the first case (Fig. 5a), birth weight  $x$  has a direct effect on mortality  $y$ , while smoking  $t$  has not. It is also possible to consider a second case where birth weight  $x$ , and possibly smoking  $t$ , have a direct effect on mortality  $y$  (Fig. 5b). To increase the plausibility of this scenario, some known common causes of lower birth weight and mortality, denoted  $z$ , can be added (Fig. 5c). In this third case, Hernández-Díaz et al. (2006) claims that the variables  $z$  might induce an association between smoking and mortality, conditional on birth weight  $x$ . Lastly, a fourth situation that combines the second and the third can be considered (Fig. 5d).

Here, instead of focusing on newborn mortality (which is an unbalanced variable, with less than 0.5% mortality rate), we consider  $y = 1$  (non-natural delivery). As can

<sup>1</sup> [https://www.cdc.gov/nchs/data\\_access/Vitalstatsonline.htm](https://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm).





**Fig. 5.** Directed acyclic graphs for the birth weight paradox, when  $y$  is the mortality indicator,  $x$  the birth weight and  $t$  a smoking indicator.  $z$  denotes some possible common causes of infant death, from Hernández-Díaz et al. (2006).

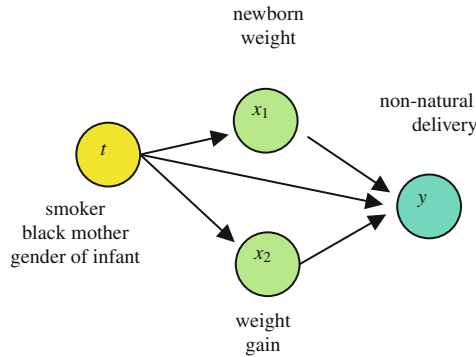
be seen in Table 2, about a third of all deliveries can be considered as “un-natural” (or “complicated”, involving a least a C-section). Among possible explanatory variables, we consider the weight of the newborn infant  $x_1$  and the weight gain of the mother  $x_2$ . Conditional densities, of  $\mathbf{x} = [x_1 \ x_2]$  given  $y$  can be visualized in Fig. 8. To illustrate various techniques based on optimal transport, we will consider  $\text{CATE}(\mathbf{x})$ ,

$$\tau(\mathbf{x}) = \text{CATE}(\mathbf{x}) = \mathbb{P}[Y_{T \leftarrow 1}^* = 1 | \mathbf{X} = \mathbf{x}] - \mathbb{P}[Y_{T \leftarrow 0}^* = 1 | \mathbf{X} = \mathbf{x}],$$

for several possible “treatment”  $t$ , that can be visualized in Fig. 6, with either a smoker indicator (for the mother) or a variable indicating whether the newborn is a boy or not. However, emphasis will be placed on a variable indicating whether the mother is Black (Afro-American) or not. Conditional densities of  $\mathbf{x}$  given  $t$  can be visualized in Fig. 9. In a nutshell, we want to address the following questions “*what would have been the probability of a non-natural delivery for a baby of weight  $x_1$  whose mother gained weight  $x_2$  during pregnancy, if the mother had been Afro-American?*” or “*if the mother had been smoking?*”

**Table 2.** Statistics about the variable of interest  $y$ , indicating a non-natural delivery, and two explanatory variables, the weight of the newborn child ( $x_1$ ) and the weight gain of the mother ( $x_2$ ), on top; and statistics about the “treatment” considered at the bottom.

	Variable of interest			
	$y = 0$ (natural)		$y = 1$ (non-natural)	
$n$ number of observations	2,221,522 (65.70%)		1,159,776 (34.30%)	
$x_1$ weight of newborn	average 3,299 g.		average 3,231 g.	
$x_2$ weight gain of mother	average 30.02 lbs.		average 31.16 lbs.	
	“Treatment”			
	$t = 0$		$t = 1$	
Afro-American variable	non-Black	2,980,387 (88.14%)	Black	400,911 (11.86%)
smoker variable	non-smoker	2,959,847 (91.54%)	smoker	273,685 (8.46%)
sex variable	baby boy	1,730,837 (51.18%)	baby girl	1,650,461 (48.82%)



**Fig. 6.** Directed acyclic graphs to explain non-natural deliveries, when  $y = \mathbf{1}$  (non-natural delivery),  $x$  is either the birth weight of the infant ( $x_1$ ), or the weight gain of the pregnant mother ( $x_2$ ), and  $t$  is either a smoker indicator (for the mother), or an indicator that the mother is Black (Afro-American), or that the baby is a boy.

### 3 Quantile Based Matching

In this section, we consider the simple case where  $x$  is univariate. This allows us to introduce properties that will be extended more formally in higher dimension in the next section. Following the example of Sect. 2.4, we will propose some techniques to generate a counterfactual version of  $(x, y, t = 0)$ , or  $(x, y_{T \leftarrow 0}^*)$ , that will be  $(x_{T \leftarrow 1}, y_{T \leftarrow 1}^*)$ . In Sect. 3.1, we will discuss classical matching techniques, used to match each point in  $(y_i, x_i, t_i = 0)$  –in the control group– with another one in  $(y_j, x_j, t_j = 1)$  –in the treated group– when the two groups have the same size. In Sect. 3.2, we will suggest an optimal matching algorithm, to associate individual  $i$  (in the control group) to  $j$  (in the treated

group), or  $j_i^*$ . Then, in Sect. 3.3, we will discuss the case where the two groups have different sizes, that will be called optimal “coupling”. In Sect. 3.4, we will define an estimator, the *mutatis mutandis* CATE,  $\widehat{m}_1(\widehat{\mathcal{F}}(x)) - \widehat{m}_0(x)$ , where  $\widehat{\mathcal{F}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ , with  $\widehat{F}_0$  and  $\widehat{F}_1$  denoting the empirical distribution functions of  $x$  conditional on  $t = 0$  and  $t = 1$ , respectively. Thus, we will use quantiles to optimal “transport”  $x$ ’s from the control group to the treated group, formally through the  $\mathcal{F}$  mapping. Finally, in Sect. 3.5, we will illustrate this on the probability that a non-natural baby delivery occurs.

### 3.1 Classical Matching Techniques

To estimate the average treatment effect  $\tau = \mathbb{E}[Y_{T \leftarrow 1}^* - Y_{T \leftarrow 0}^*]$ , a standard technique is to consider matching techniques to match each point in  $(y_i, x_i, t_i = 0)$  or  $(y_i^{(0)}, x_i^{(0)})$  with another one in  $(y_j, x_j, t_j = 1)$ , or  $(y_j^{(1)}, x_j^{(1)})$ . In this coupling approach, we assume that there are  $n$  treated and  $n$  non-treated individuals. A treated individual  $i$  ( $t_i = 1$ ) is matched to someone in the non-treated group ( $t_j = 0$ ) that is close enough for some distance on the set of covariates  $\mathcal{X}$ ,  $j_i^* = \operatorname{argmin}_{j:t_j=0} \{d(x_i^{(0)}, x_j^{(1)})\}$ , so that

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^n (y_i^{(1)} - y_{j_i^*}^{(0)}) = \frac{1}{n} \sum_{i=1}^n y_{j_i^*}^{(1)} - \frac{1}{n} \sum_{i=1}^n y_i^{(0)} = \bar{y}^{(1)} - \bar{y}^{(0)},$$

since we simply consider a re-ordering of the treated population. But interestingly, that approach provides a counterfactual version of  $(x_i, y_i)$  in the treated population,  $(x_{j_i^*}, y_{j_i^*})$ . The algorithm performing such a matching would be Algorithm 1.

---

#### Algorithm 1. Counterfactual matching – “1:1 nearest neighbor matching” (classical)

---

```

 $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ 
function COUNTERFACTUAL1( $\mathcal{D}$ )
     $\mathcal{D}_0 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 0$  (size  $n$ ) shuffled, with indices  $i$ 
     $\mathcal{D}_1 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 1$  (size  $n$ ), with indices  $j$ 
    for  $i = 1, 2, \dots, n$  do
         $j_i^* = \operatorname{argmin}_{j:t_j=1} \{d(\mathbf{x}_i, \mathbf{x}_j)\}$  in  $\mathcal{D}_1$ ,
         $L_i \leftarrow (i, j_i^*, y_{j_i^*}^{(1)} - y_i^{(0)})$ 
        remove observation  $j_i^*$  from  $\mathcal{D}_1$ 
    end for
    return matrix  $L$  ( $n \times 3$ , with  $L = (L_i)$ )
end function
    
```

---

This algorithm, introduced by Rubin (1973), is described in Stuart (2010) under the name “1:1 nearest neighbor matching”, and properties are discussed in Ho et al. (2007) or Dehejia and Wahba (1999) that focuses on the problem of not removing selected observations (also called “Greedy Matching”).

Quite naturally, it is possible to define some local version of the previous quantity using weights or some  $k$  nearest neighbors approach, to derive an estimate of the CATE  $\widehat{\tau}(x)$ , as in Algorithm 2

$$\widehat{\tau}(x) \propto \sum_{i=1}^n \omega_i(x) (y_{j_i^*}^{(1)} - y_i^{(0)}),$$

where weight  $\omega_i(x)$  are all the higher that  $x_i$  is close to  $x$ , either based on a  $k$ -nearest neighbors approach ( $\omega_i(x) = \mathbf{1}(i \in V_x^k)$ , as in Algorithm 2) or based on a kernel approach ( $\omega_i(x) = K(|x - x_i|)$  for some kernel  $K$ ).

Unfortunately, that matching mechanism can be very sensitive to the initial permutation: individuals picked first will have a counterfactual in the treated group close to them, but it might not be the case for the individuals picked last. In the next section, we will consider some optimal matching among individuals in the two populations.

---

**Algorithm 2.** Estimate SCATE (classical, with  $k$ -NN)

---

dataset  $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ ,

**function** SCATE1( $\mathcal{D}, k, \mathbf{x}$ )

$L \leftarrow \text{COUNTERFACTUAL1}(\mathcal{D})$

$V_x^k \leftarrow$  list of  $k$  nearest neighbors of  $\mathbf{x}$ ;  $t_i$ 's in  $\mathcal{D}_0$  close to  $\mathbf{x}$

**for**  $i \in V_x^k$  **do**

$d_i \leftarrow L \$ d(i)$

**end for**

**return**  $\frac{1}{k} \sum_{i \in V_x^k} d_i$

**end function**

---

### 3.2 Optimal Matching

The matching procedure described previously is characterized by some  $n \times n$  permutation matrix,  $P$ , with entries in  $\{0, 1\}$ , satisfying  $\mathbb{P}\mathbf{1}_n = \mathbf{1}_n$  and  $\mathbb{P}^{\star\top}\mathbf{1}_n = \mathbf{1}_n$ , see Brualdi (2006). Hence, there is a permutation  $\sigma$  of  $\{1, \dots, n\}$  such that  $j_i^* = \sigma(i)$ , and  $P$  is the matrix associated with  $\sigma$  (that satisfies  $\mathbf{e}_i P = \mathbf{e}_{\sigma(i)}$ , where  $\mathbf{e}_i$ 's denote the standard basis vector, i.e., a row vector of length  $n$  with 1 in the  $i$ -th position and 0 in every other position). It is possible to seek an ‘‘optimal’’ permutation: if  $C$  is the  $n \times n$  matrix that quantifies the distance between individuals in the two groups,  $C_{i,j} = d(x_i^{(0)}, x_j^{(1)}) = \delta(x_i^{(0)} - x_j^{(1)})$ , the optimal matching is solution of

$$\min_{P \in \mathcal{P}} \langle P, C \rangle = \min_{P \in \mathcal{P}} \sum_{i,j} P_{i,j} C_{i,j},$$

where  $\mathcal{P}$  is the set of permutation matrices, and  $\langle \cdot, \cdot \rangle$  is the Frobenius dot-product. This is also called Kantorovich’s optimal transport problem, from Kantorovich (1942). If  $\delta$  is (strictly) convex –as is the standard Euclidean distance– it can be proven that this optimal transport problem has a simple solution. Instead of using  $(y_i^{(0)}, x_i^{(0)})$ , let  $r_i^{(0)}$

denote the rank of  $x_i^{(0)}$  in  $\{x_1^{(0)}, \dots, x_n^{(0)}\}$ . Similarly, let  $r_i^{(1)}$  denote the rank of  $x_i^{(1)}$  in the treated dataset  $\{x_1^{(1)}, \dots, x_n^{(1)}\}$ . The procedure then becomes simply a matching based on ranks, in the sense that  $J_i^*$  satisfies  $r_{J_i^*}^{(1)} = r_i^{(0)}$ , as discussed in Chap. 2 of Santambrogio (2015). Since ranks are defined on  $\{1, 2, \dots, n\}$ , vectors  $\mathbf{r}^{(0)}$  and  $\mathbf{r}^{(1)}$  correspond to two permutations of  $\{1, 2, \dots, n\}$ , that we can denote  $\sigma_0$  and  $\sigma_1$ , respectively. The optimal coupling is based on permutation  $\sigma = \sigma_1 \circ \sigma_0^{-1}$  in the sense that  $x_i^{(0)}$  is associated to  $x_{\sigma(i)}^{(1)}$ . If the  $\mathbf{x}^{(0)}$ 's and the  $\mathbf{x}^{(1)}$ 's are sorted, then  $P = \mathbb{I}_n$ , i.e.,  $x_i^{(0)}$  is coupled with  $x_i^{(1)}$ . Or, if  $\widehat{F}_0$  and  $\widehat{F}_1$  are the cumulative distribution functions associated with sample  $\mathbf{x}^{(0)}$  and  $\mathbf{x}^{(1)}$ , we can see that if  $u \in (0, 1)$  is such that  $\widehat{F}_0^{-1}(u) = x_i^{(0)}$ , then  $\widehat{F}_1^{-1}(u) = x_i^{(1)}$ , with the exact same  $i$ .

### 3.3 Optimal Coupling

The previous procedure can be extended in the case where the two groups do not necessarily have the same size. If the two groups  $(x_i, t_i = 0)$  and  $(x_j, t_j = 1)$  have different sizes, namely  $n_0$  and  $n_1$ , respectively, it is possible to define some matching using weights, and weighted mean of individuals in the two groups.

In a very general setting, if  $\mathbf{a}_0 \in \mathbb{R}_+^{n_0}$  and  $\mathbf{a}_1 \in \mathbb{R}_+^{n_1}$  satisfy  $\mathbf{a}_0^\top \mathbf{1}_{n_0} = \mathbf{a}_1^\top \mathbf{1}_{n_1}$  (identical sums), define

$$U(\mathbf{a}_0, \mathbf{a}_1) = \left\{ M \in \mathbb{R}_+^{n_0 \times n_1} : M \mathbf{1}_{n_1} = \mathbf{a}_0 \text{ and } M^\top \mathbf{1}_{n_0} = \mathbf{a}_1 \right\}.$$

This set of matrices is a convex polytope (see Brualdi (2006)). The optimal coupling is matrix  $P^*$  solution of

$$\min_{P \in U(\mathbf{a}_0, \mathbf{a}_1)} \{ \langle C, P \rangle \},$$

which is solved using linear programming, by casting matrix  $P \in \mathbb{R}_+^{n_0 \times n_1}$  as a vector  $\mathbf{p} \in \mathbb{R}_+^{n_0 n_1}$  such that  $\mathbf{p}_{i+n(j-1)} = P_{i,j}$ , and similarly for the cost matrix  $C$ . The constraint  $P \in U(\mathbf{a}_0, \mathbf{a}_1)$  becomes equivalently

$$\begin{pmatrix} \mathbf{1}_{n_0}^\top \otimes \mathbb{I}_{n_1} \\ \mathbb{I}_{n_0} \otimes \mathbf{1}_{n_1}^\top \end{pmatrix} \mathbf{p} = A \mathbf{p} = (\mathbf{a}_0, \mathbf{a}_1)^\top = \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \end{pmatrix},$$

where  $A$  is some  $(n_0 + n_1) \times (n_0 n_1)$  matrix. The optimal matching problem is then simply

$$\min \left\{ \mathbf{c}^\top \mathbf{p} \right\} \text{ subject to } A \mathbf{p} = (\mathbf{a}_0, \mathbf{a}_1)^\top.$$

In our case, let  $U_{n_0, n_1}$  denote  $U(\mathbf{1}_0, \frac{n_0}{n_1} \mathbf{1}_1)$

$$P^* \in \operatorname{argmin}_{P \in U_{n_0, n_1}} \langle P, C \rangle \text{ ou } \operatorname{argmin}_{P \in U_{n_0, n_1}} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} P_{i,j} C_{i,j}. \quad (2)$$

One can notice that this matrix optimisation problem does not depend on the dimension of space, so it will easily be extended to the case where  $x$  is multivariate. Nevertheless, in the univariate setting, this approach can be related to quantile functions.

### 3.4 From Optimal Matching to CATE

Let  $F_0$  and  $F_1$  denote the two conditional distributions of  $X$ , an absolutely continuous variable, in the control group ( $t = 0$ ) and in the treatment group ( $t = 1$ ), respectively. Then the optimal matching between the two groups is based on transformation  $\mathcal{T} : x_0 \mapsto x_1 = F_1^{-1} \circ F_0(x_0)$ . From the probability integral transform property: if  $X_0 \sim F_0$ , then  $F_0(X_0)$  is uniform on the unit interval  $[0, 1]$ , and then  $X_1 = \mathcal{T}(X_0) \sim F_1$ .

**Lemma 1.** *If  $X_0 \sim F_0$ , then  $X_1 = \mathcal{T}(X_0) \sim F_1$ , where  $\mathcal{T} : x_0 \mapsto x_1 = F_1^{-1} \circ F_0(x_0)$ .*

**Definition 2.** The *mutatis mutandis* quantile-based CATE is

$$\text{QCATE}(u) = \mathbb{E}[Y_{T \leftarrow 1}^* | X = F_1^{-1}(u)] - \mathbb{E}[Y_{T \leftarrow 0}^* | X = F_0^{-1}(u)], \quad (3)$$

where  $F_t$  is the cumulative distribution function of  $X$ , conditional on  $T = t$ , or

$$\text{CATE}(x) = \mathbb{E}[Y_{T \leftarrow 1}^* | X = \mathcal{T}(x)] - \mathbb{E}[Y_{T \leftarrow 0}^* | X = x], \quad \mathcal{T} = F_1^{-1} \circ F_0 \quad (4)$$

where  $x$  is considered with respect to the control group.

Thus,  $\text{CATE}(x) = \text{QCATE}(F_0(x))$ .

**Definition 3.** Consider two models,  $\widehat{m}_0(x)$  and  $\widehat{m}_1(x)$ , that estimate, respectively,  $\mathbb{E}[Y|X = x, T = 0]$  and  $\mathbb{E}[Y|X = x, T = 1]$ . A natural estimator of the *mutatis mutandis* CATE is

$$\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{T}}(x)) - \widehat{m}_0(x)$$

where  $\widehat{\mathcal{T}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ ,  $\widehat{F}_0$  with  $\widehat{F}_1$  denoting the empirical distribution functions of  $x$  conditional on  $t = 0$  and  $t = 1$ , respectively.

Note that a simple parametric transformation can be obtained, based on the assumption that  $X$  conditional on  $T$  is Gaussian. More precisely, if  $X_1 \stackrel{\mathcal{L}}{=} X|t = 1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$  and  $X_0 \stackrel{\mathcal{L}}{=} X|t = 0 \sim \mathcal{N}(\mu_0, \sigma_0)$ ,

$$\mu_1 + \sigma_1 \cdot \frac{X_0 - \mu_0}{\sigma_0} \stackrel{\mathcal{L}}{=} X_1$$

**Definition 4.** Consider two models,  $\widehat{m}_0(x)$  and  $\widehat{m}_1(x)$ , that estimate respectively  $\mathbb{E}[Y|X = x, T = 0]$  and  $\mathbb{E}[Y|X = x, T = 0]$ . A Gaussian estimator of the *mutatis mutandis* CATE is

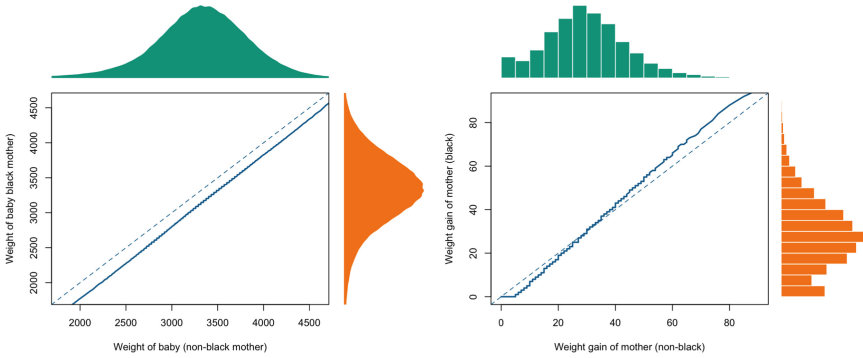
$$\text{SCATE}_{\mathcal{N}}(x) = \widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$$

where  $\widehat{\mathcal{T}}_{\mathcal{N}}(x) = \bar{x}_1 + s_1 s_0^{-1}(x - \bar{x}_0)$ ,  $\bar{x}_0$  and  $\bar{x}_1$  being respectively the averages of  $x$  in the two sub-populations, and  $s_0$  and  $s_1$  the sample standard deviations.

An algorithm to compute that estimator is Algorithm 6 (in higher dimension).

### 3.5 Application to Non-natural Deliveries

In Fig. 7, we can visualize  $x \mapsto \widehat{\mathcal{T}}(x)$  when  $x$  is either the weight of the newborn infant on the left, or the weight gain of the mother on the right, when  $t$  indicates whether the mother is Black or not. The  $x$ -axis is the value of  $x$  in the control group ( $t = 0$ ) and the  $y$ -axis is the value of  $x$  in the treated group ( $t = 1$ ). On the left, observe that  $x \mapsto \widehat{\mathcal{T}}(x)$  is almost linear, parallel to the first diagonal, below. This corresponds to the fact that the distribution of  $X$  conditional on  $T = 0$  and  $T = 1$  are similar, up to a translation (same standard deviation but different mean if a Gaussian transport  $\widehat{\mathcal{T}}_{\mathcal{N}}$  was considered). On the right,  $x \mapsto \widehat{\mathcal{T}}(x)$  is single-crossing the first diagonal. This corresponds to the fact that the distribution of  $X$  conditional on  $T = 0$  and  $T = 1$  have different variances.

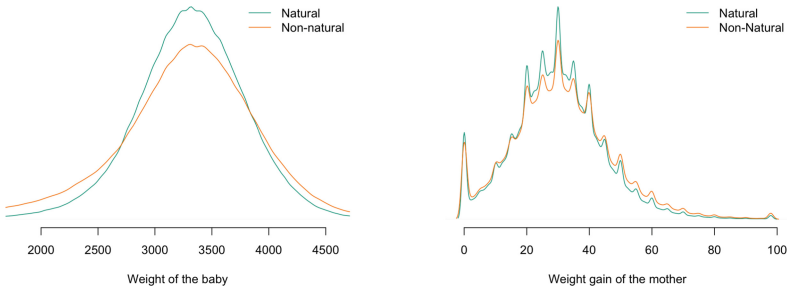


**Fig. 7.** Optimal transport (quantile based) when  $X$  is the weight of the newborn infant on the left, and the weight gain of the mother on the right, when  $T$  indicates whether the mother is Black or not in the middle. The solid line depicts the transported values while the dashed line is the identity line. See Fig. 26 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

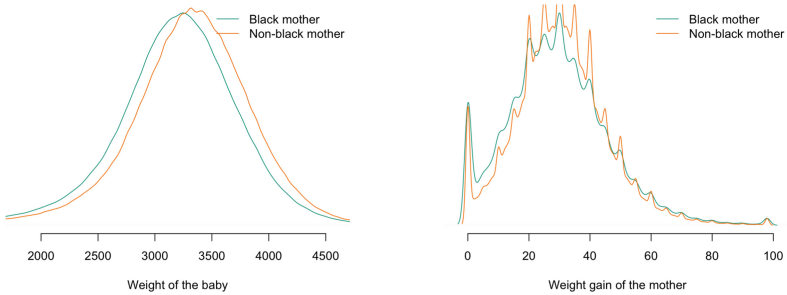
In Fig. 8, we can visualize the conditional distributions of  $x$ , when  $y = 0$  and  $y = 1$  (natural and non-natural deliveries, respectively), when  $x$  is the weight of the baby (on the left) and the weight gain of the mother (on the right). In Fig. 9, we can visualize the conditional distributions of  $x$ , when  $y = 0$  and  $y = 1$ , when  $t = 0$  and  $t = 1$ , where  $t$  denotes whether the mother is Afro-American or not.

In Fig. 7, we can visualize the empirical optimal coupling function  $\widehat{\mathcal{T}} : x_0 \mapsto x_1 = \widehat{F}_1^{-1} \circ \widehat{F}_0(x_0)$ , where  $\widehat{F}_0$  and  $\widehat{F}_1$  denote the empirical distribution functions of  $x$  conditional on  $t = 0$  and  $t = 1$ , respectively.

In Figs. 10 and 11, we can visualize  $\widehat{m}_0(x)$  and  $\widehat{m}_1(\widehat{\mathcal{T}}(x))$  on the left, when  $t$  indicates whether the mother is Afro-American or not, when  $x$  the weight of the newborn infant in Fig. 10 and when  $x$  is the weight gain of the mother in Fig. 11. On the right, we can visualize  $x \mapsto \text{CATE}(x) = \widehat{m}_1(\widehat{\mathcal{T}}(x)) - \widehat{m}_0(x)$  as a function of  $x$ . The light curve in the back is  $\widehat{m}_1(x) - \widehat{m}_0(x)$ . Numerical values are given in Table 3 when  $x$  is the weight of the newborn, and Table 4 when  $x$  is the weight gain of the pregnant mother.



**Fig. 8.** Distribution of the weight of the newborn infant (in grams) on the left and distribution of the weight gain of the mother on the right, conditional on the delivery mode,  $Y = \mathbf{1}$  (non-natural delivery).



**Fig. 9.** Distribution of the weight of the newborn infant (in grams) on the left and distribution of the weight gain of the mother on the right, whether the mother is Black or not. See Fig. 17 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

For instance, a baby weighting 2500g (7.46% quantile in the non-Black population) corresponds to a baby weighting 2301g if the mother had been Black. The probability to have a non-natural delivery has then an additional 5.5% compared with non-Black mother, using the GAM-SCATE approach. Using a Gaussian transport, the counterfactual in the Black population is a 2297g baby, and the probability to have a non-natural delivery has then an additional 5.60% compared with a non-Black mother, using the  $\text{GAM-SCATE}_{\mathcal{N}}$  approach. Similarly, a baby weighting 3500g (64.13% quantile in the non-Black population) corresponds to a baby weighting 3375g had the mother been Black (about 3.6% less). The probability to have a non-natural delivery has then an additional 4.42% compared with non-Black mother, using the GAM-SCATE approach. Using a Gaussian transport, estimates are similar.

In Fig. 12, as previously,  $\hat{m}_0(x)$  and  $\hat{m}_1(\widehat{\mathcal{F}}_{\mathcal{N}}(x))$  can be visualized on the left, when  $t$  indicates whether the mother is Afro-American or not, and when  $x$  is the gain weight of the mother. On the right, we can visualize  $x \mapsto \text{CATE}(x) = \hat{m}_1(\widehat{\mathcal{F}}(x)) - \hat{m}_0(x)$  as a function of  $x$ . Numerical values are given in Table 3 when  $x$  is the weight of the newborn, and Table 4 when  $x$  is the weight gain of the pregnant mother.



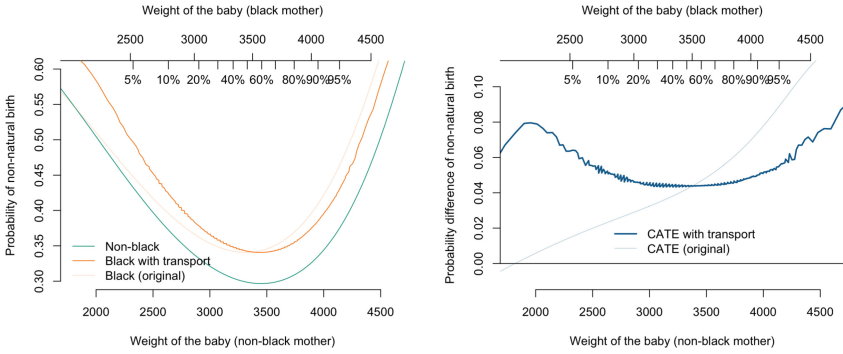
**Table 3.** Estimation of the conditional average treatment (CATE), on the probability to have a non-natural birth ( $y$ ), as a function of the weight of the baby ( $x$ , in g.), when the mother is Afro-American. Several weights  $x$  are considered, from 2 to 4.5kg.  $u$  is the probability associated with  $x$ , in the baseline population ( $t = 0$ ).  $\text{CATE}_0$  is simply the difference  $\widehat{m}_1(x) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs.  $\widehat{\mathcal{F}}(x)$  is the quantile based transport function ( $\widehat{\mathcal{F}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ ), while  $\widehat{\mathcal{F}}_{\mathcal{N}}(x)$  is the Gaussian one. Thus,  $\text{SCATE}(x)$  is the *mutatis mutandis* CATE  $\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{F}}(x)) - \widehat{m}_0(x)$ , while  $\text{SCATE}_{\mathcal{N}}(x) = \widehat{m}_1(\widehat{\mathcal{F}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs. Finally, the last estimate is obtained when  $\widehat{m}_0$  and  $\widehat{m}_1$  are simple local averages, using kernels. See Table 5 in Appendix 5.2 for similar table when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

$t$ : mother is Afro-American						
$x$ (newborn's weight)	2000	2500	3000	3500	4000	4500
$u$	2.67%	7.46%	25.13%	64.13%	91.73%	98.87%
$\text{CATE}_0(x)$ (GAM)	0.58%	1.99%	3.24%	4.86%	7.78%	11.70%
$\widehat{\mathcal{F}}(x)$	1595	2301	2863	3375	3890	4415
$\text{SCATE}(x)$ (GAM)	7.94%	5.53%	4.53%	4.42%	5.16%	7.46%
$\widehat{\mathcal{F}}_{\mathcal{N}}(x)$	1758	2297	2836	3376	3915	4455
$\text{SCATE}_{\mathcal{N}}(x)$ (GAM)	5.15%	5.60%	4.82%	4.42%	5.71%	9.41%
$\text{SCATE}_{\mathcal{N}}(x)$ (kernel)	6.98%	6.64%	4.34%	4.53%	5.34%	7.13%

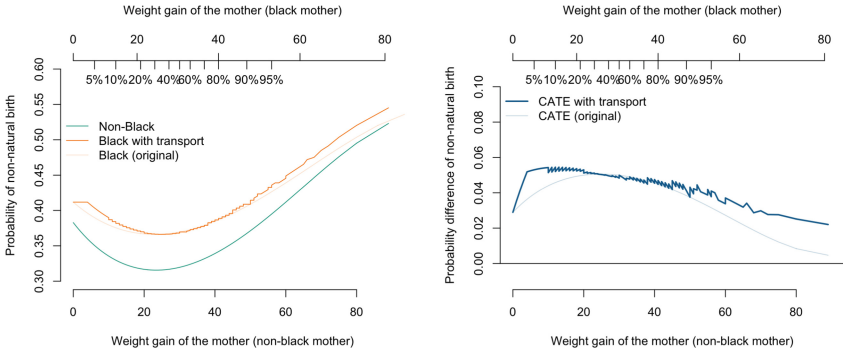
**Table 4.** Estimation of the conditional average treatment (CATE), on the probability to have a non-natural birth ( $y$ ), as a function of the weight gain of the mother ( $x$ , in lbs), when the mother is Afro-American. Several weight gains  $x$  are considered, from 5 to 55lbs.  $u$  is the probability associated with  $x$ , in the baseline population ( $t = 0$ ).  $\text{CATE}_0$  is simply the difference  $\widehat{m}_1(x) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs.  $\widehat{\mathcal{F}}(x)$  is the quantile based transport function ( $\widehat{\mathcal{F}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ ), while  $\widehat{\mathcal{F}}_{\mathcal{N}}(x)$  is the Gaussian one. Thus,  $\text{SCATE}(x)$  is the *mutatis mutandis* CATE  $\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{F}}(x)) - \widehat{m}_0(x)$ , while  $\text{SCATE}_{\mathcal{N}}(x) = \widehat{m}_1(\widehat{\mathcal{F}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs. Finally, the last estimate is obtained when  $\widehat{m}_0$  and  $\widehat{m}_1$  are simple local averages, using kernels. See Table 6 in Appendix 5.2 for similar table when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

$t$ : mother is Afro-American						
$x$ (weight gain of the mother)	5	15	25	35	45	55
$u$	4.57%	14.34%	37.15%	66.81%	86.34%	94.94
$\text{CATE}_0(x)$ (GAM)	3.79%	4.79%	5.06%	4.82%	4.18%	3.26%
$\widehat{\mathcal{F}}(x)$	1	12	24	35	47	58
$\text{CATE}(x)$ (GAM)	5.25%	5.25%	5.04%	4.82%	4.69%	4.19%
$\widehat{\mathcal{F}}_{\mathcal{N}}(x)$	1	12	23	34	46	57
$\text{CATE}_{\mathcal{N}}(x)$ (GAM)	5.22%	5.21%	5.03%	4.74%	4.33%	3.78%
$\text{CATE}_{\mathcal{N}}(x)$ (kernel)	3.78%	5.49%	5.31%	4.49%	4.12%	3.61%

In Fig. 13, some local kernels are used to estimate  $\widehat{m}_0(x)$  and  $\widehat{m}_1(\widehat{\mathcal{F}}_{\mathcal{N}}(x))$  on the left. Numerical values are given in Table 3 when  $x$  is the weight of the newborn, and Table 4 when  $x$  is the weight gain of the pregnant mother.



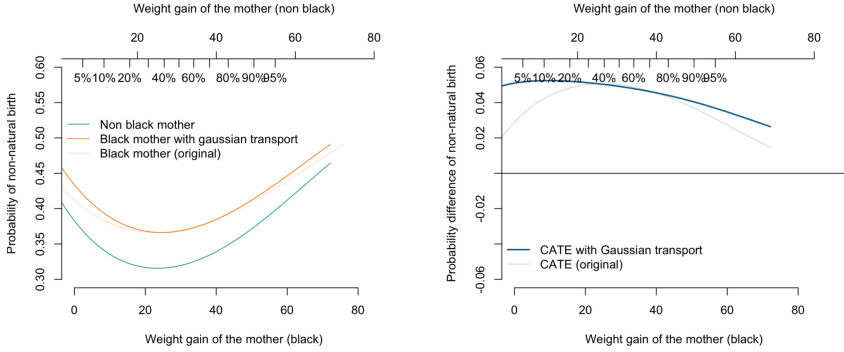
**Fig. 10.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}$  (non-natural delivery), and  $X$  is the weight of the newborn infant, respectively when  $T$  indicates whether the mother is Black or not. On the right, evolution of  $x \mapsto \text{SCATE}[Y|X = x]$ . See Fig. 18 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.



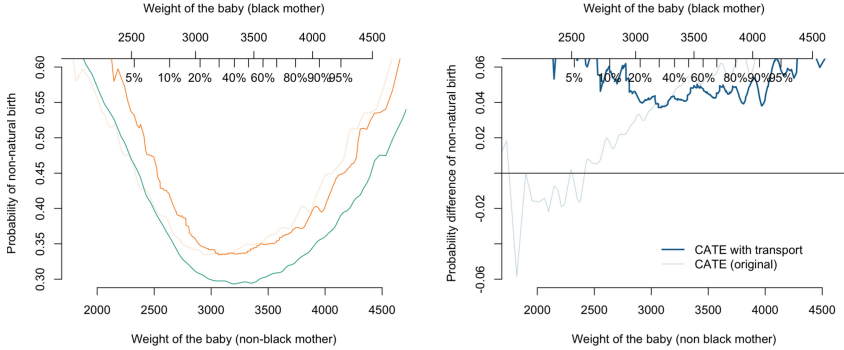
**Fig. 11.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}$  (non-natural delivery), and  $X$  is the weight gain of the mother, respectively when  $T$  indicates whether the mother is Black or not. On the right, evolution of  $x \mapsto \text{SCATE}[Y|X = x]$ . See Fig. 19 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

## 4 Optimal Transport Based Matching

In this section, we will extend what was derived in the previous section. Heuristically, optimal matching of margins components of  $\mathbf{x}$  will probably not work, and the mapping should be multivariate. We will therefore use optimal transport techniques to get a proper counterfactual of  $\mathbf{x}$ , not in the control group, but in the treated group. In Sect. 4.1, we will define properly the optimal transport problem (in any dimension). Then, in Sect. 4.2, we will describe how to optimally associate each observation  $\mathbf{x}_i$  in the control group (when  $t = 0$ ) with a single counterfactual observation  $\mathbf{x}_j$  in the treated group (when  $t = 1$ ), when two groups have the same size. This can be related to the Gaus-



**Fig. 12.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}$ (non-natural delivery), and  $X$  is the weight gain of the mother, respectively when  $T$  indicates whether the mother is Black or not. On the right, evolution of  $x \mapsto \text{SCATE}_{\mathcal{N}}[Y|X = x]$ . See Fig. 20 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not.



**Fig. 13.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a kernel based local average, when  $Y = \mathbf{1}$ (non-natural delivery), and  $X$  is the weight of the newborn infant, respectively when  $T$  indicates whether the mother is a smoker or not (on top), when the mother is Black or not in the middle, and the sex of the infant below. On the right, evolution of  $x \mapsto \text{SCATE}_{\mathcal{N}}[Y|X = x]$  with an without transport, based on a Gaussian transport. See Fig. 22 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not, or indicates the sex of the newborn.

sian SEM discussed in Sect. 2.4. In Sect. 4.3, we will present the extension when the two groups have different sizes. In Sect. 4.4, we will give an explicit formulation for  $\mathcal{S}$  when we the distribution of  $\mathbf{X}$  conditional on  $T$  is assumed to be Gaussian. The application to non-natural deliveries will finally be discussed in Sect. 4.5.

### 4.1 Optimal Transport

In the mathematical formulation of Monge (1781)’s problem, we want to push a distribution from  $\mathbb{P}_0$  to  $\mathbb{P}_1$  (distributions on  $\mathbb{R}^k$ , not necessarily in  $\mathbb{R}$  as considered in the previous section). Given  $\mathcal{S} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ , define the “push-forward” measure,

$$\mathbb{P}_1(A) = \mathcal{T}_\# \mathbb{P}_0(A) = \mathbb{P}_0(\mathcal{T}^{-1}(A)), \forall A \subset \mathbb{R}^k.$$

For instance, when  $k = 1$ , if  $F$  is the cumulative distribution of a univariate random variable  $X$  under  $\mathbb{P}$  (i.e.,  $F(x) = \mathbb{P}[X \leq x]$ ) then  $\mathbb{Q} = F_\# \mathbb{P}$  is the uniform distribution on the unit interval  $[0, 1]$  as well as  $\mathbb{Q}' = \bar{F}_\# \mathbb{P}$ , where  $\bar{F}$  is the survival function associated with  $F$  (i.e.,  $\bar{F}(x) = \mathbb{P}[X > x]$ ). Similarly, or conversely, if  $Q$  is the quantile function associated with  $F$   $-Q(u) = F^{-1}(u)$  for any  $u \in (0, 1)$ – then if  $\mathbb{P}$  is the uniform distribution on the unit interval  $[0, 1]$ ,  $\mathbb{Q} = Q_\# \mathbb{P}$  satisfies  $\mathbb{Q}[X \leq x] = Q^{-1}(x) = F(x)$ , and similarly for  $\bar{\mathbb{Q}}$  where  $\bar{\mathbb{Q}}(u) = F^{-1}(1 - u)$ .

Observe that if  $\mathbb{P}_0$  and  $\mathbb{P}_1$  have densities  $f_0$  and  $f_1$ , respectively, and if  $T$  is continuously differentiable,  $\mathbb{P}_1 = \mathcal{T}_\# \mathbb{P}_0$  is any only if  $f_0(\mathbf{x}) = f_1(\mathcal{T}(\mathbf{x})) \cdot |\det \nabla \mathcal{T}(\mathbf{x})|$ , for all  $\mathbf{x}$ . This non-linear function is a special case of the so-called Monge-Ampère partial differential equations.

An optimal transport  $\mathcal{T}^*$  (in Brenier’s sense, from Brenier (1991), see Villani (2009) or Galichon (2016)) from  $\mathbb{P}_0$  towards  $\mathbb{P}_1$  will be solution of

$$\mathcal{T}^* \in \operatorname{arginf}_{\mathcal{T}: \mathbb{P}_0 = \mathbb{P}_1} \left\{ \int_{\mathbb{R}^k} \|\mathbf{x} - \mathcal{T}(\mathbf{x})\|^2 d\mathbb{P}_0(\mathbf{x}) \right\},$$

for a quadratic cost, or more generally,

$$\mathcal{T}^* \in \operatorname{arginf}_{\mathcal{T}: \mathbb{P}_0 = \mathbb{P}_1} \left\{ \int_{\mathbb{R}^k} \gamma(\mathbf{x}, \mathcal{T}(\mathbf{x})) d\mathbb{P}_0(\mathbf{x}) \right\},$$

for some cost function  $\gamma: \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}_+$ .

If  $k = 1$ , and if the cost function  $\gamma$  can be written  $\gamma(x, y) = h(|x - y|)$  for some strictly convex and positive function  $h$ , then  $T^*$  is an increasing function, and more precisely, if  $F_0(x) = \mathbb{P}_0[X \leq x]$  and  $F_1(x) = \mathbb{P}_1[X \leq x]$ , with  $F_0$  absolutely continuous, then  $\mathcal{T}^*(x) = F_1^{-1} \circ F_0(x)$  satisfies  $\mathcal{T}^*_\# \mathbb{P}_0 = \mathbb{P}_1$  (since  $F_1(x) = F_0(T^{*-1}(x))$ ) and  $\mathcal{T}^*$  is optimal. the quadratic cost function (when  $h(x) = x^2$ ) is a particular case. The case where  $h$  is concave was discussed in McCann (1999).

In higher dimension, for a quadratic cost, one can prove (see Villani (2003; 2009) or Galichon (2016)) that  $\mathcal{T}^* = \nabla \psi$  where  $\psi$  is a convex function.

## 4.2 Empirical Version of Optimal Matching

This transport can be seen as a matching between individuals in the two groups, both of size  $n$ ,  $(\mathbf{x}_i, t_i = 0)$  and  $(\mathbf{x}_j, t_j = 1)$ , instead of two distributions  $\mathbb{P}_0$  and  $\mathbb{P}_1$ . If  $C$  is a  $n \times n$  matrix that quantifies the distance between individuals in the two groups,  $C_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$ , the optimal matching is solution of

$$\min_{P \in \mathcal{P}} \langle P, C \rangle = \min_{P \in \mathcal{P}} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j},$$

where  $\mathcal{P}$  is the set of permutation matrices, and  $\langle \cdot, \cdot \rangle$  is the Frobenius dot-product. This is also called Kantorovich’s optimal transport problem, from Kantorovich (1942). Interestingly, there are some algorithms that can be used to find that optimal coupling, or matching, which can, in turn, be used to get a counterfactual for all individuals in each group.

### 4.3 Empirical Version of Optimal Coupling

If the two groups ( $\mathbf{x}_i, t_i = 0$ ) and ( $\mathbf{x}_j, t_j = 1$ ) have different sizes, namely  $n_0$  and  $n_1$ , respectively, it is possible to define some matching using weights. In the coupling case, described previously,  $P$  was some  $n \times n$  permutation matrix. But here, as in Sect. 3.3 some  $n_0 \times n_1$  matrices will be involved, and similar problems are considered

$$\min_{P \in U(\mathbf{a}_0, \mathbf{a}_1)} \langle P, C \rangle = \min_{P \in U(\mathbf{a}_0, \mathbf{a}_1)} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} P_{i,j} C_{i,j}. \quad (5)$$

And again, assuming Gaussian distributions for  $\mathbf{X}$  conditional on  $T$  will provide an explicit simple transport formula that can be used to get an estimation of the *mutatis mutandis* CATE. This algorithm is given by Algorithm 4, used to compute the Average Treatment Effect.

---

#### Algorithm 3. Counterfactual matching, with optimal matching

---

```

 $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ 
function COUNTERFACTUAL2( $\mathcal{D}$ )
     $\mathcal{D}_0 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 0$  (size  $n_0$ ), with indices  $i$ 
     $\mathcal{D}_1 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 1$  (size  $n_1$ ), with indices  $j$ 
     $C \leftarrow$  matrix  $n_0 \times n_1$ ,  $C_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$  between points in  $\mathcal{D}_0$  and  $\mathcal{D}_1$ 
     $P^* \leftarrow$  solution of Problem (2)
    return matrix  $P^*$  ( $n_0 \times n_1$ )
end function
    
```

---



---

#### Algorithm 4. Estimate SCATE (optimal matching based)

---

```

dataset  $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ ,
function SCATE2( $\mathcal{D}, k, \mathbf{x}$ )
     $P \leftarrow$  COUNTERFACTUAL2( $\mathcal{D}$ )
     $V_{\mathbf{x}}^k \leftarrow$  list of  $k$  nearest neighbors of  $\mathbf{x}_i$ 's in  $\mathcal{D}_0$  close to  $\mathbf{x}$ 
    return  $\frac{1}{k} \sum_{i \in V_{\mathbf{x}}^k} y_i^1 - P_i^\top \mathbf{y}^0$ 
end function
    
```

---

### 4.4 Counterfactuals for Gaussian Covariates

In the general case, there are no simple construction and interpretation of the optimal mapping  $\mathcal{T}^*$ , as the one we had in the univariate case, based on quantiles. If it is possible, following Marc (2021), to define multivariate quantiles (and therefore to extend concepts defined in Sect. 3.4). But here, we will simply consider the multivariate Gaussian case. Suppose that  $\mathbf{X}|t = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  and  $\mathbf{X}|t = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ . There is an

explicit expression for the optimal transport, which is simply an affine map (see Villani (2003) for more details). In the univariate case,  $x_1 = \mathcal{T}_{\mathcal{N}}^*(x_0) = \mu_1 + \frac{\sigma_1}{\sigma_0}(x_0 - \mu_0)$ , while in the multivariate case, an analogous expression can be derived:

$$\mathbf{x}_1 = \mathcal{T}_{\mathcal{N}}^*(\mathbf{x}_0) = \boldsymbol{\mu}_1 + \mathbf{A}(\mathbf{x}_0 - \boldsymbol{\mu}_0),$$

where  $\mathbf{A}$  is a symmetric positive matrix that satisfies  $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A} = \boldsymbol{\Sigma}_1$ , which has a unique solution given by  $\mathbf{A} = \boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2})^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$ , where  $\mathbf{M}^{1/2}$  is the square root of the square (symmetric) positive matrix  $\mathbf{M}$  based on the Schur decomposition ( $\mathbf{M}^{1/2}$  is a positive symmetric matrix), as described in Higham (2008).

**Definition 5.** Consider two models,  $\widehat{m}_0(\mathbf{x})$  and  $\widehat{m}_1(\mathbf{x})$ , that estimate, respectively,  $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0]$  and  $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1]$ . A Gaussian estimator of the *mutatis mutandis* CATE is

$$\text{SCATE}_{\mathcal{N}}(\mathbf{x}) = \widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(\mathbf{x})) - \widehat{m}_0(\mathbf{x}),$$

where  $\widehat{\mathcal{T}}_{\mathcal{N}}(\mathbf{x}) = \bar{\mathbf{x}}_1 + \widehat{\mathbf{A}}(\mathbf{x} - \bar{\mathbf{x}}_0)$ , with  $\bar{\mathbf{x}}_0$  and  $\bar{\mathbf{x}}_1$  being, respectively, the averages of  $\mathbf{x}$  in the two sub-populations, and  $\widehat{\mathbf{A}} = \widehat{\boldsymbol{\Sigma}}_0^{-1/2}(\widehat{\boldsymbol{\Sigma}}_0^{1/2}\widehat{\boldsymbol{\Sigma}}_1\widehat{\boldsymbol{\Sigma}}_0^{1/2})^{1/2}\widehat{\boldsymbol{\Sigma}}_0^{-1/2}$  where  $\widehat{\boldsymbol{\Sigma}}_0$  and  $\widehat{\boldsymbol{\Sigma}}_1$  denote the sample variance.

The algorithm to compute that estimate is Algorithm 6.

We should probably stress here that, in the very general case, we should transport *only* endogenous variables  $\mathbf{x}^m$  (or mediators) and not exogenous ones  $\mathbf{x}^c$  (or colliders), as discussed in Sect. 2.1 (and Fig. 2).

---

#### Algorithm 5. Optimal Gaussian Transport

---

dataset  $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ ,

**function** TGAUSSIAN( $\mathcal{D}$ )

$\mathcal{D}_0 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 0$

$\mathcal{D}_1 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 1$

    estimate moments of  $\mathbf{x}$ 's  $\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_0$  and  $\hat{\boldsymbol{\Sigma}}_1$ , in  $\mathcal{D}_0$  and  $\mathcal{D}_1$

$\hat{\mathbf{A}} \leftarrow \hat{\boldsymbol{\Sigma}}_0^{-1/2}(\hat{\boldsymbol{\Sigma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_0^{1/2})^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1/2}$

**function** T( $\mathbf{x}$ )

**return**  $\hat{\boldsymbol{\mu}}_1 + \hat{\mathbf{A}}(\mathbf{x} - \hat{\boldsymbol{\mu}}_0)$ ,

**end function**

**return function** T

**end function**

---

**Algorithm 6.** Parametric Estimate  $\text{SCATE}_{\mathcal{N}}$  (Gaussian transport)

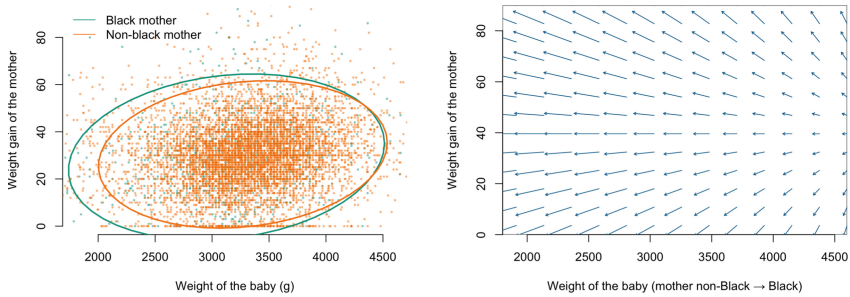
---

```

dataset  $\mathcal{D} \leftarrow \{(y_i, \mathbf{x}_i, t_i)\}$ ,
 $\mathcal{D}_0 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 0$ 
 $\mathcal{D}_1 \leftarrow$  subset of  $\mathcal{D}$  when  $t = 1$ 
 $\hat{m}_0 \leftarrow$  model to predict  $y$  based on  $\mathbf{x}$ , trained on  $\mathcal{D}_0$ 
 $\hat{m}_1 \leftarrow$  model to predict  $y$  based on  $\mathbf{x}$ , trained on  $\mathcal{D}_1$ 
 $T \leftarrow \text{TGAUSSIAN}(\mathcal{D})$ 
function SCATE3( $\hat{m}_0, \hat{m}_1, T, \mathbf{x}$ )
    return  $\hat{m}_1(T(\mathbf{x})) - \hat{m}_0(\mathbf{x})$ 
end function

```

---

**4.5 Application to Non-natural Deliveries**

**Fig. 14.** Joint distributions of  $\mathbf{X}$  (weight of the newborn infant and weight gain of the mother), conditional on the treatment  $T$ , when  $T$  indicates whether the mother is Black or not on the left. On the right, vector field associated with optimal Gaussian transport, in dimension two (weight of the newborn infant and weight gain of the mother). Some numerical values are given in Table 7. On the right, the origin of the arrow is  $\mathbf{x}$  in the control group (non-Black pregnant mother) and the arrowhead is  $\widehat{\mathcal{T}}_{\mathcal{N}}(\mathbf{x})$  in the treated group (Black pregnant mother). See Fig. 25 in Appendix 5.2 for similar graphs when  $T$  indicates whether the pregnant mother is a smoker or not.

The left-hand side of Fig. 14 displays a scatter plot of  $\mathbf{x} = (x_1, x_2)$ , where  $x_1$  represents the weight of the newborn infant while  $x_2$  shows the weight gain of the mother, conditional on the treatment  $T$ , when  $T$  indicates whether the mother is Black or not (see Fig. 25 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not). The ellipses are the iso-density curves under a Gaussian assumption, such that 95% of the points lie in the ellipse. The right-hand side of Fig. 14, shows  $\widehat{\mathcal{T}}_{\mathcal{N}}$  on the same frame,  $\mathbf{x} = (x_1, x_2)$ , with, respectively, the weight of the newborn infant on the  $x$ -axis and weight gain of the mother on the  $y$ -axis. The origin of an arrow corresponds to  $\mathbf{x} = (x_1, x_2)$ , while its end corresponds to  $\widehat{\mathcal{T}}_{\mathcal{N}}(\mathbf{x})$ . Note that all the arrows point to the left. Regardless of the weight of the mother, had the latter been Black, the weight of the newborn would have been lower. Nevertheless, the length of the arrows varies according to the weight of the newborn. For infants whose weight

is relatively high, for example for  $x_1$  close to 4500g, had the mother been Black, the newborn's weight would have been almost the same. For newborns whose weight  $x_1$  is much lower than 4500g, had the mother been Black, the baby's weight would have been much smaller. Some numerical values are given in Table 7 in Appendix 5.2. For instance, if we consider a non-Black mother with a baby weighting 2584g, who gained 10.8lbs, the counterfactual is a Black mother with a baby weighting 2392g, who gained 7.6lbs.

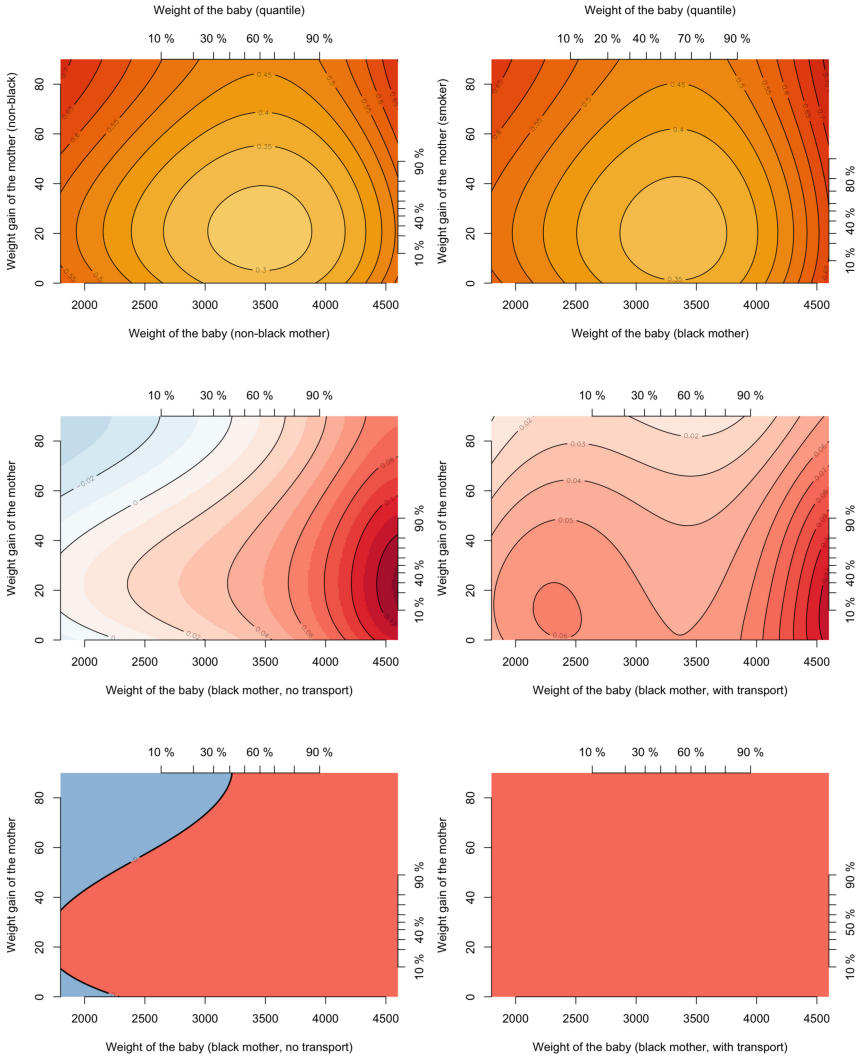
The top panel of Figs. 15 shows the level curves of  $\widehat{m}_0 : \mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0]$  (left-hand side) and  $\widehat{m}_1 : \mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1]$  (right-hand side), when the treatment  $T$  indicates whether a mother is Black or not, estimated with logistic GAM models (cubic splines). The middle-level panel displays curves of the *ceteris paribus*  $\mathbf{x} \mapsto \text{CATE}[\mathbf{x}]$  without any transport (on the left), and  $\mathbf{x} \mapsto \text{SCATE}[\mathbf{x}]$  *mutatis mutandis* (on the right). Lastly, the bottom panel shows a positive/negative distinction for the conditional average treatment effect (positive is red, negative is blue). Figure 16 provides different results using more knots in the cubic splines. We can observe that all mothers are more likely to get a non-natural delivery would they be Black, whatever the weight of the baby (the *ceteris paribus* approach would suggest that mothers with small babies, below 2.5kg would be less likely to get a non-natural delivery if they were Black).

As briefly discussed earlier, optimal matching or coupling in high dimension can be computationally intensive, since matrices  $n_0 \times n_1$  are involved. For instance, when  $t$  is the sex of the newborn, the cost matrix is a matrix with 3,000 billion entries. Thus, it is quite natural to consider sub-sampling techniques (since our dataset is quite large). For convenience, we can use optimal matching on groups of size  $n$ , and study the robustness of estimated, as a function of  $n$ . Some simulations are mentioned in the Appendix.

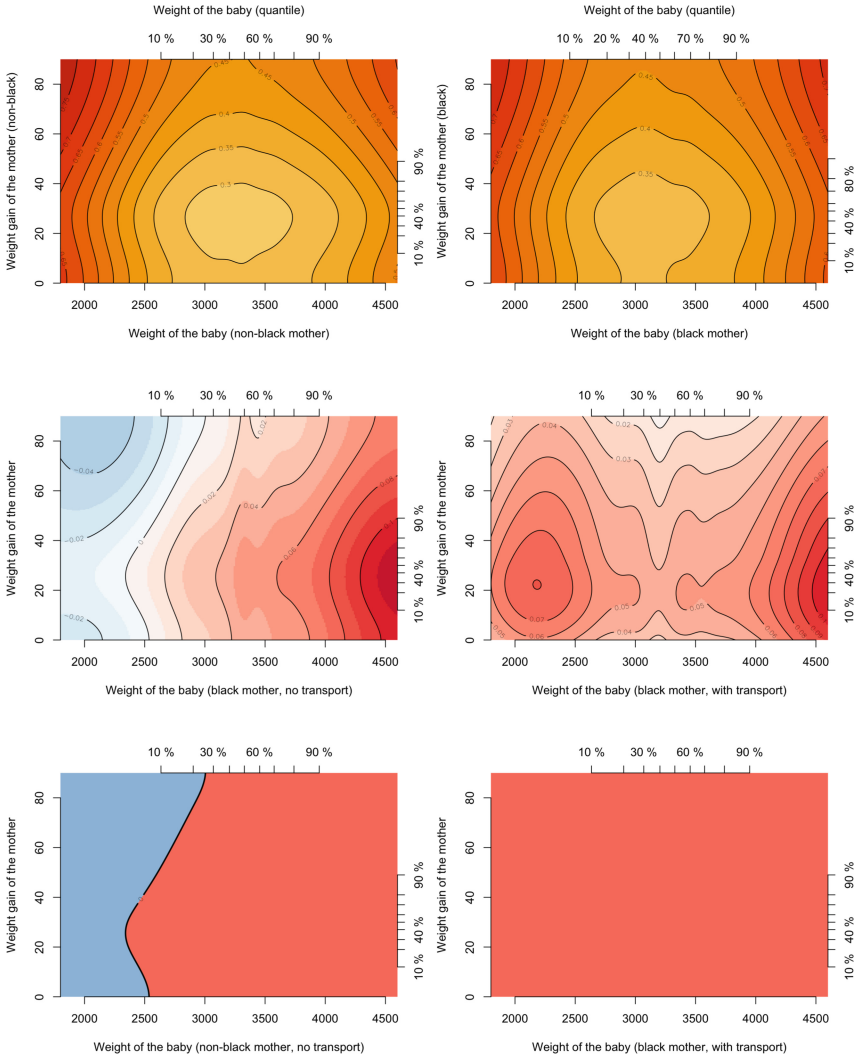
**Acknowledgments.** Arthur Charpentier acknowledges the financial support of the AXA Research Fund through the joint research initiative *use and value of unusual data in actuarial science*, as well as NSERC grant 2019-07077.

Emmanuel Flachaire and Ewen Gallic acknowledge the financial support of the French National Research Agency Grant ANR-17-EURE-0020, the Excellence Initiative of Aix Marseille University – A\*MIDEX.





**Fig. 15.** On top, contours of  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0]$  and  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1]$  when  $T$  indicates whether a mother is Afro-American or not, estimated with logistic GAM models (cubic splines). In the middle, contours of the *ceteris paribus*  $\mathbf{x} \mapsto \text{CATE}[\mathbf{x}]$  without any transport on the left, and  $\mathbf{x} \mapsto \text{SCATE}[\mathbf{x}]$  *mutatis mutandis* on the right. At the bottom, positive/negative distinction for the conditional average treatment effect. See Fig. 28 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not.



**Fig. 16.** On top, contours of  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X}=\mathbf{x}, T=0]$  and  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X}=\mathbf{x}, T=1]$  when  $T$  indicates whether a mother is Afro-American or not, estimated with logistic GAM models (cubic splines, **with more knots and degrees of freedom**). In the middle, contours of the *ceteris paribus*  $\mathbf{x} \mapsto \text{CATE}[\mathbf{x}]$  without any transport on the left, and  $\mathbf{x} \mapsto \text{SCATE}[\mathbf{x}]$  *mutatis mutandis* on the right. At the bottom, positive/negative distinction for the conditional average treatment effect. See Fig. 28 in Appendix 5.2 for similar graphs when  $T$  indicates whether the mother is a smoker or not.

## 5 Appendix

### 5.1 Estimation of CATE in a Gaussian framework

With a correlation  $r$  (in the simulations, we considered  $r = 0.4$ ), consider the following SEM,

$$\begin{cases} T = \mathbf{1}(\varepsilon_t < 0), \\ X_1^m = \varepsilon_{1m} + (T = 1) \cdot (2 + 0.2\varepsilon_{1m}) \\ X_2^m = r\varepsilon_{1m} + \sqrt{1-r^2}\varepsilon_{2m} - 0.2(T = 1) \cdot (r\varepsilon_{1m} + \sqrt{1-r^2}\varepsilon_{2m}), \\ X^c = \varepsilon_c \\ Y = 2 + T + X_1^m - X_2^m + X^c + \varepsilon_y \end{cases}$$

where  $\varepsilon$ 's are independent  $\mathcal{N}(0, 1)$  variables. The two interventions yield

$$\begin{array}{cc} do(T = 0) & do(T = 1) \\ \left\{ \begin{array}{l} T \leftarrow 0 \\ X_1^m = \varepsilon_{1m} \\ X_2^m = r\varepsilon_{1m} + \sqrt{1-r^2}\varepsilon_{2m} \\ X^c = \varepsilon_c \\ Y_{T \leftarrow 0} = 2 + X_1^m - X_2^m + X^c + \varepsilon_y \end{array} \right. & \left\{ \begin{array}{l} T \leftarrow 1 \\ X_1^m = 2 + 1.2\varepsilon'_{1m} \\ X_2^m = 0.8r\varepsilon'_{1m} + \sqrt{1-r^2}\varepsilon'_{2m} \\ X^c = \varepsilon'_c \\ Y_{T \leftarrow 1} = 3 + X_1^m - X_2^m + X^c + \varepsilon'_y \end{array} \right. \end{array}$$

and if we consider  $do(T = 0)$ , when  $X_1^m = x_1$ , we have

$$\left\{ \begin{array}{l} T \leftarrow 0 \\ X_1^m = x_1^m \\ X_2^m = rx_1^m + \sqrt{1-r^2}\varepsilon_{2m} \\ X^c = \varepsilon_c \\ Y_{T \leftarrow 0} = 2 + x_1^m - (rx_1^m + \sqrt{1-r^2}\varepsilon_{2m}) + \varepsilon_c + \varepsilon_y \end{array} \right.$$

while if we consider  $do(T = 1)$ , when  $X_1^m = x'_1$

$$\left\{ \begin{array}{l} T \leftarrow 1 \\ X_1^m = x'_1 \\ X_2^m = 0.8r(x'_1 - 2)/1.2 + \sqrt{1-r^2}\varepsilon'_{2m} \\ X^c = \varepsilon'_c \\ Y_{T \leftarrow 1} = 3 + x_1^m - (0.8r(x'_1 - 2)/1.2 + \sqrt{1-r^2}\varepsilon'_{2m}) + \varepsilon'_c + \varepsilon'_y \end{array} \right.$$

so that

$$\left\{ \begin{array}{l} Y_{T \leftarrow 0} = 2 + (1-r)x'_1 + \sqrt{1-r^2}\varepsilon_{2m} + \varepsilon_c + \varepsilon_y \\ Y_{T \leftarrow 1} = 3 + 1.6r/1.2 + (1-0.8r/1.2)x'_1 - \sqrt{1-r^2}\varepsilon'_{2m} + \varepsilon'_c + \varepsilon'_y \end{array} \right.$$

Since  $X_1^m = \varepsilon'_{1m}$  when  $T \leftarrow 0$ , while  $X_1^m = 2 + 1.2\varepsilon'_{1m}$  when  $T \leftarrow 1$ , it is legitimate to assume that if  $x_{1,T \leftarrow 0}^m = x_1$ , then  $x_{1,T \leftarrow 1}^m = 2 + 1.2x_{1,T \leftarrow 0}^m$ , in a *mutatis mutandis* approach,

$$\left\{ \begin{array}{l} Y_{T \leftarrow 0} = 2 + (1-r)x_1 + \sqrt{1-r^2}\varepsilon_{2m} + \varepsilon_c + \varepsilon_y \\ Y_{T \leftarrow 1} = 3 + 1.6r/1.2 + (1-0.8r/1.2)(2 + 1.2x_1) - \sqrt{1-r^2}\varepsilon'_{2m} + \varepsilon'_c + \varepsilon'_y \end{array} \right.$$

Thus,

$$ATE = \mathbb{E}[Y_{T \leftarrow 1} - Y_{T \leftarrow 0}] = 3$$

while the *mutatis mutandis* CATE is

$$CATE(x_1) = \mathbb{E}[Y_{T \leftarrow 1} | x_{1, T \leftarrow 1}^m] - \mathbb{E}[Y_{T \leftarrow 0} | x_{1, T \leftarrow 0}^m = x_1]$$

that is

$$CATE(x_1) = [3 + 1.6r/1.2 + (1 - 0.8r/1.2)(2 + 1.2x_1)] - [2 + (1 - r)x_1]$$

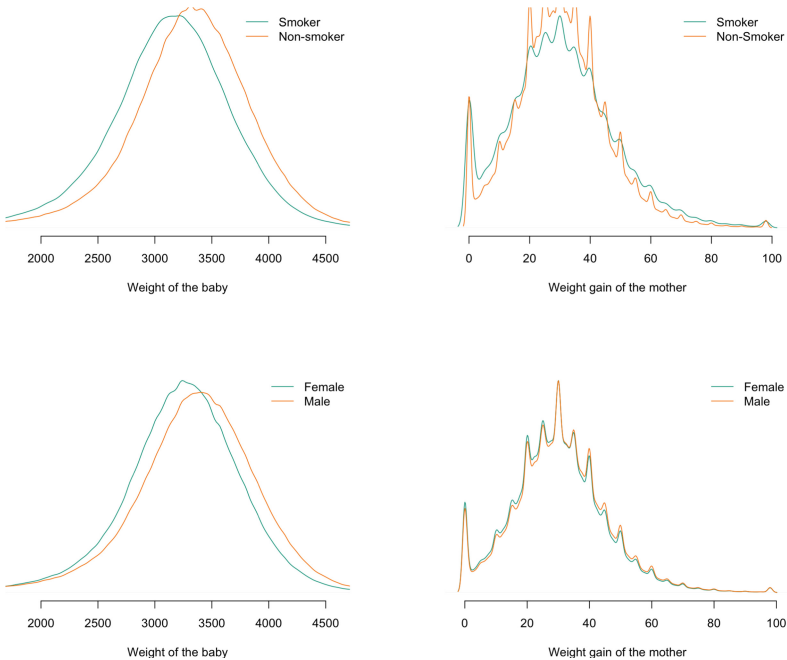
i.e.,

$$CATE(x_1) = 3 + 0.2(1 + r)x_1$$

that is linear in  $x_1^m$ , with slope  $0.2(1 + r)$  in this *mutatis mutandis* case.

### 5.2 Additional Applications (Smoker and Sex of Newborn)

In this section, similar graphs to the one presented earlier are produced (Figs. 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 and 30).



**Fig. 17.** Distribution of the weight of the newborn infant (in grams) on the left and distribution of the weight gain of the mother on the right, when the mother is a smoker or not on top, and depending on the sex of the newborn infant at the bottom.

**Table 5.** Estimation of the conditional average treatment (CATE), on the probability to have a non-natural birth ( $y$ ), as a function of the weight of the baby ( $x$ , in g.), for different “treatments” ( $t$ ): when the mother is a smoker, and when the baby is a boy. Several weights  $x$  are considered, from 2 to 4.5 kg.  $u$  is the probability associated with  $x$ , in the baseline population ( $t = 0$ ).  $\text{CATE}_0$  is simply the difference  $\widehat{m}_1(x) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs.  $\widehat{\mathcal{F}}(x)$  is the quantile based transport function ( $\widehat{\mathcal{F}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ ), while  $\widehat{\mathcal{F}}_{\mathcal{N}}(x)$  is the Gaussian one. Thus,  $\text{SCATE}(x)$  is the *mutatis mutandis* CATE  $\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{F}}(x)) - \widehat{m}_0(x)$ , while  $\text{SCATE}_{\mathcal{N}}(x) = \widehat{m}_1(\widehat{\mathcal{F}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs. Finally, the last estimate is obtained when  $\widehat{m}_0$  and  $\widehat{m}_1$  are simple local averages, using kernels.

$t$ : mother smoker						
$x$ (newborn’s weight)	2000	2500	3000	3500	4000	4500
$u$	2.75%	7.44%	24.99%	64.14%	91.75%	98.86%
$\text{CATE}_0(x)$ (GAM)	-4.41%	-2.55%	-0.69%	0.79%	1.97%	2.50%
$\widehat{\mathcal{F}}(x)$	1775	2280	2802	3317	3830	4337
$\text{SCATE}(x)$ (GAM)	-0.08%	1.15%	1.15%	0.50%	-1.07%	-4.39%
$\widehat{\mathcal{F}}_{\mathcal{N}}(x)$	1786	2295	2805	3314	3824	4333
$\text{SCATE}_{\mathcal{N}}(x)$ (GAM)	-0.28%	0.88%	1.12%	0.50%	-1.15%	-4.53%
$\text{SCATE}_{\mathcal{N}}(x)$ (kernel)	-0.80%	0.24%	1.72%	0.15%	-1.75%	-2.78%
$t$ : sex of the newborn						
$x$ (newborn’s weight)	2000	2500	3000	3500	4000	4500
$u$	2.79%	7.26%	23.00%	60.32%	90.04%	98.55%
$\text{CATE}_0(x)$ (GAM)	-0.24%	-1.66%	-2.24%	-2.00%	-0.77%	2.38%
$\widehat{\mathcal{F}}(x)$	1960	2438	2892	3374	3856	4338
$\text{SCATE}(x)$ (GAM)	0.71%	-0.38%	-0.96%	-2.04%	-3.38%	-5.14%
$\widehat{\mathcal{F}}_{\mathcal{N}}(x)$	1947	2424	2901	3377	3854	4331
$\text{SCATE}_{\mathcal{N}}(x)$ (GAM)	1.02%	-0.08%	-1.07%	-2.05%	-3.41%	-5.43%
$\text{SCATE}_{\mathcal{N}}(x)$ (kernel)	2.06%	-0.27%	-1.02%	-2.30%	-3.53%	-5.10%

In Fig. 29,  $\text{SCATE}(\mathbf{x})$  is estimated for various values of  $\mathbf{x}$  ( $\mathbf{x} = (2500, 60)$  on top,  $\mathbf{x} = (4200, 60)$  in the middle and  $\mathbf{x} = (2500, 20)$  at the bottom), depending on sample size  $n$ . The solid line is the average value, that is quite stable, but, as expected, the confidence interval is quite large when  $n$  is small. In Fig. 30, we can visualize the distribution of  $\widehat{m}_0(\mathbf{x})$  on top,  $\widehat{m}_1(\widehat{\mathcal{F}}(\mathbf{x}))$  in the middle, and  $\text{SCATE}(\widehat{\mathcal{F}}(\mathbf{x})) = \widehat{m}_1(\widehat{\mathcal{F}}(\mathbf{x})) - \widehat{m}_0(\mathbf{x})$  at the bottom, estimated on bootstrapped samples of size  $n = 20,000$ . On the left,  $\mathbf{x} = (2500, 20)$  and on the right  $\mathbf{x} = (2500, 60)$ . The two densities are based on the fact that two GAM models are considered, with more or less knots. Confidence intervals are obtained by bootstrap.

**Table 6.** Estimation of the conditional average treatment (CATE), on the probability to have a non-natural birth ( $y$ ), as a function of the weight gain of the mother ( $x$ , in lbs), for different “treatments” ( $t$ ): when the mother is a smoker, and when the baby is a boy. Several weight gains  $x$  are considered, from 5 to 55 lbs.  $u$  is the probability associated with  $x$ , in the baseline population ( $t = 0$ ).  $\text{CATE}_0(x)$  is simply the difference  $\widehat{m}_1(x) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs.  $\widehat{\mathcal{T}}(x)$  is the quantile based transport function ( $\widehat{\mathcal{T}}(x) = \widehat{F}_1^{-1} \circ \widehat{F}_0(x)$ ), while  $\widehat{\mathcal{T}}_{\mathcal{N}}(x)$  is the Gaussian one. Thus,  $\text{SCATE}(x)$  is the *mutatis mutandis* CATE  $\text{SCATE}(x) = \widehat{m}_1(\widehat{\mathcal{T}}(x)) - \widehat{m}_0(x)$ , while  $\text{SCATE}_{\mathcal{N}}(x) = \widehat{m}_1(\widehat{\mathcal{T}}_{\mathcal{N}}(x)) - \widehat{m}_0(x)$ , where both  $\widehat{m}_0$  and  $\widehat{m}_1$  are GAMs. Finally, the last estimate is obtained when  $\widehat{m}_0$  and  $\widehat{m}_1$  are simple local averages, using kernels.

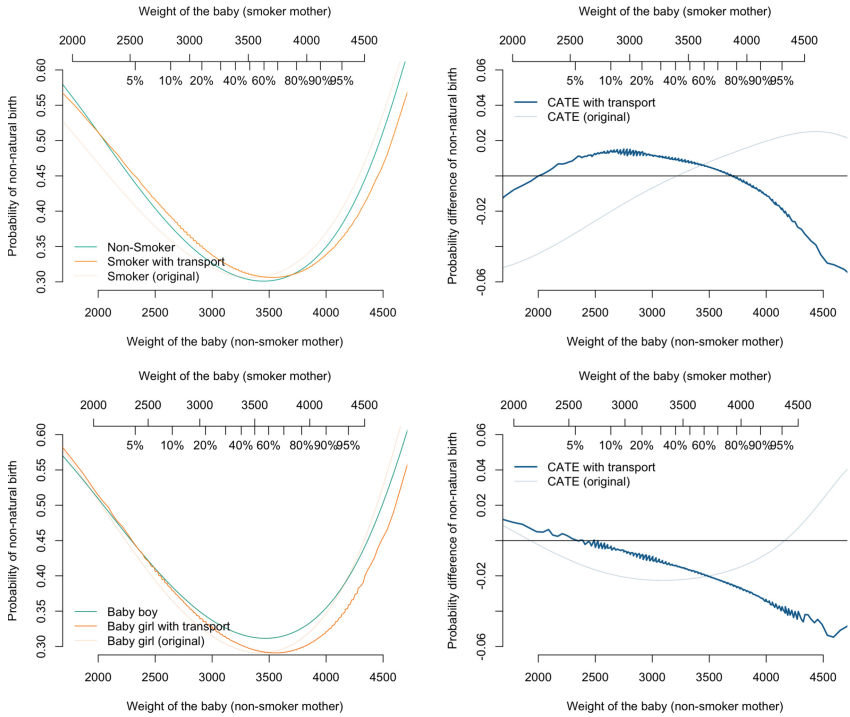
<i>t</i> : mother smoker						
<i>x</i> (weight gain of the mother)	5	15	25	35	45	55
<i>u</i>	4.61%	14.50%	37.49%	67.12%	86.54%	95.05%
$\text{CATE}_0(x)$ (GAM)	0.06%	0.84%	0.69%	-0.08%	-1.26%	-2.59%
$\widehat{\mathcal{T}}(x)$	1	13	25	37	49	60
$\text{CATE}(x)$ (GAM)	1.60%	1.21%	0.69%	0.14%	-0.38%	-1.08%
$\widehat{\mathcal{T}}_{\mathcal{N}}(x)$	1	13	24	36	48	59
$\text{CATE}_{\mathcal{N}}(x)$ (GAM)	1.63%	1.29%	0.71%	0.01%	-0.71%	-1.33%
$\text{CATE}_{\mathcal{N}}(x)$ (kernel)	0.31%	1.03%	0.98%	0.29%	-1.30%	-1.08%

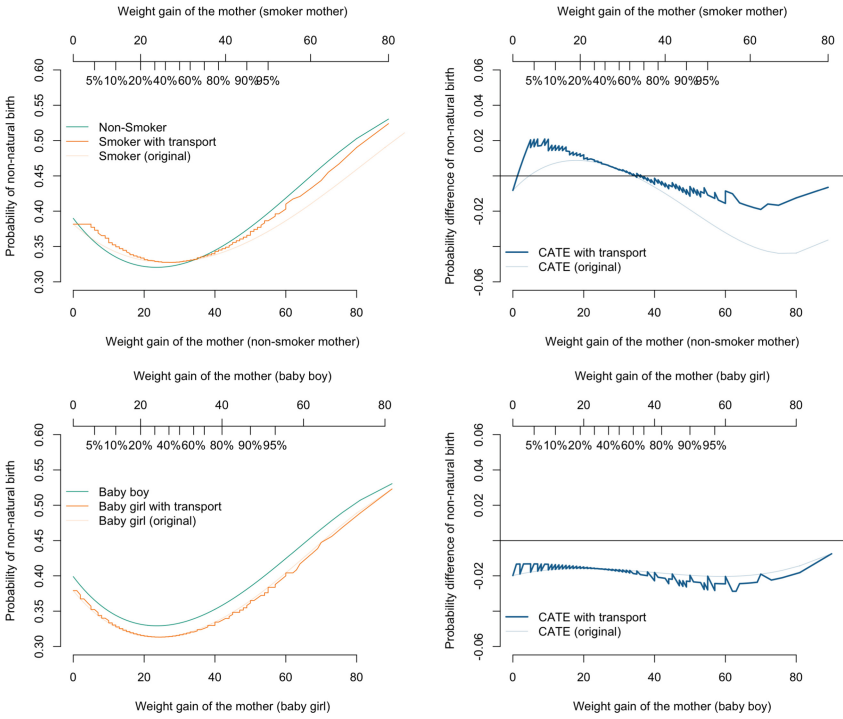
<i>t</i> : sex of the newborn						
<i>x</i> (weight gain of the mother)	5	15	25	35	45	55
<i>u</i>	4.68%	14.40%	36.73%	65.81%	85.58%	94.55%
$\text{CATE}_0(x)$ (GAM)	-1.79%	-1.60%	-1.60%	-1.73%	-1.90%	-2.02%
$\widehat{\mathcal{T}}(x)$	5	15	25	35	45	55
$\text{CATE}(x)$ (GAM)	-1.79%	-1.60%	-1.60%	-1.73%	-1.90%	-2.02%
$\widehat{\mathcal{T}}_{\mathcal{N}}(x)$	4	14	24	34	44	54
$\text{CATE}_{\mathcal{N}}(x)$ (GAM)	-1.52%	-1.47%	-1.61%	-1.87%	-2.17%	-2.41%
$\text{CATE}_{\mathcal{N}}(x)$ (kernel)	-1.58%	-1.37%	-1.66%	-1.79%	-2.22%	-2.88%

**Table 7.** Bivariate optimal transport,  $\mathbf{x} \mapsto \mathcal{T}_{\mathcal{N}}(\mathbf{x})$ , for the three treatments, for four different individuals  $\mathbf{x}$  in the control group.

<i>t</i> : mother is smoker				<i>t</i> : mother is Afro-American				<i>t</i> : Sex of the newborn			
<i>t</i> = 0		<i>t</i> = 1		<i>t</i> = 0		<i>t</i> = 1		<i>t</i> = 0		<i>t</i> = 1	
non-smoker		smoker		non-Black		Black		boy		girl	
$\mathbf{x} = (x_1, x_2)$		$\mathcal{T}_{\mathcal{N}}(\mathbf{x})$		$\mathbf{x} = (x_1, x_2)$		$\mathcal{T}_{\mathcal{N}}(\mathbf{x})$		$\mathbf{x} = (x_1, x_2)$		$\mathcal{T}_{\mathcal{N}}(\mathbf{x})$	
weight	gain	weight	gain	weight	gain	weight	gain	weight	gain	weight	gain
2584	10.8	2353.1	7.6	2584	10.8	2392.5	7.6	2584	10.8	2513.4	10.2
2584	46.8	2414.8	49.5	2584	46.8	2382.0	47.6	2584	46.8	2493.0	45.9
4152	10.8	3938.1	7.9	4152	10.8	4086.1	7.5	4152	10.8	4012.4	10.1
4152	46.8	3999.8	49.8	4152	46.8	4075.6	47.6	4152	46.8	3992.0	45.8



**Fig. 18.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T-t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}(\text{non-natural delivery})$ , and  $X$  is the weight of the newborn infant, respectively when  $T$  indicates whether the mother is a smoker or not (on top), or whether the newborn infant is a boy at the bottom. On the right, evolution of  $x \mapsto \text{SCATE}[Y|X = x]$ .

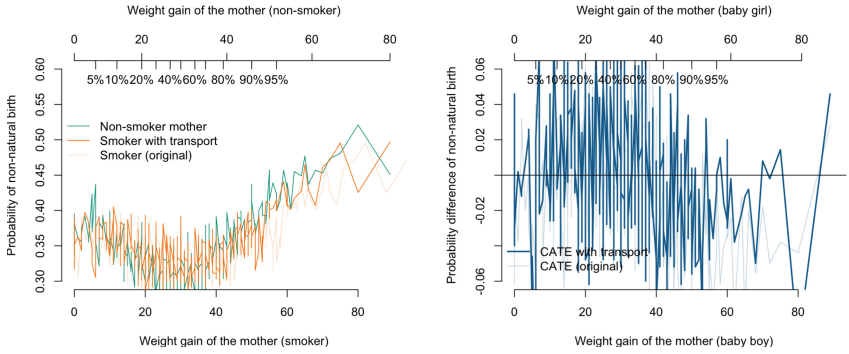


**Fig. 19.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}$  (non-natural delivery), and  $X$  is the weight gain of the mother, respectively when  $T$  indicates whether the mother is a smoker or not on top, or whether the newborn infant is a girl or not, below. On the right, evolution of  $x \mapsto \text{SCATE}[Y|X = x]$ .

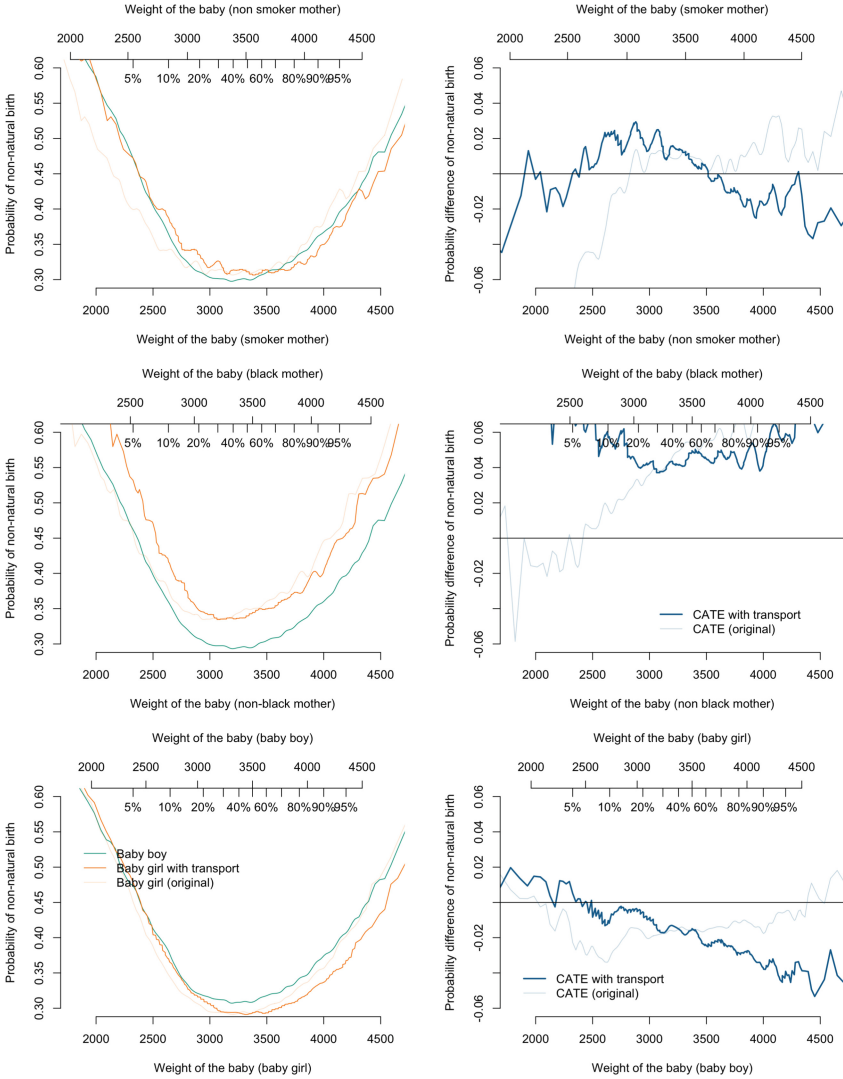




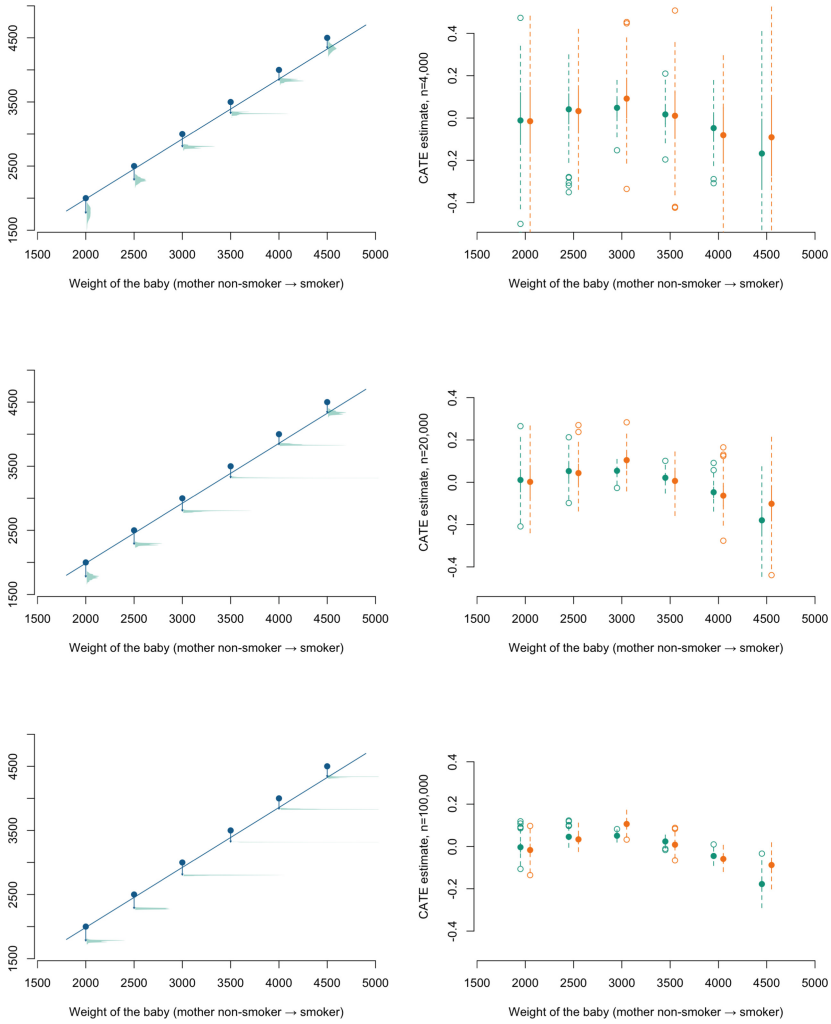
**Fig. 20.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a logistic GAM model, when  $Y = \mathbf{1}$ (non-natural delivery), and  $X$  is the weight gain of the mother, respectively when  $T$  indicates whether the mother is a smoker or not. On the right, evolution of  $x \mapsto \text{SCATE}_{\mathcal{N}}[Y|X = x]$ .



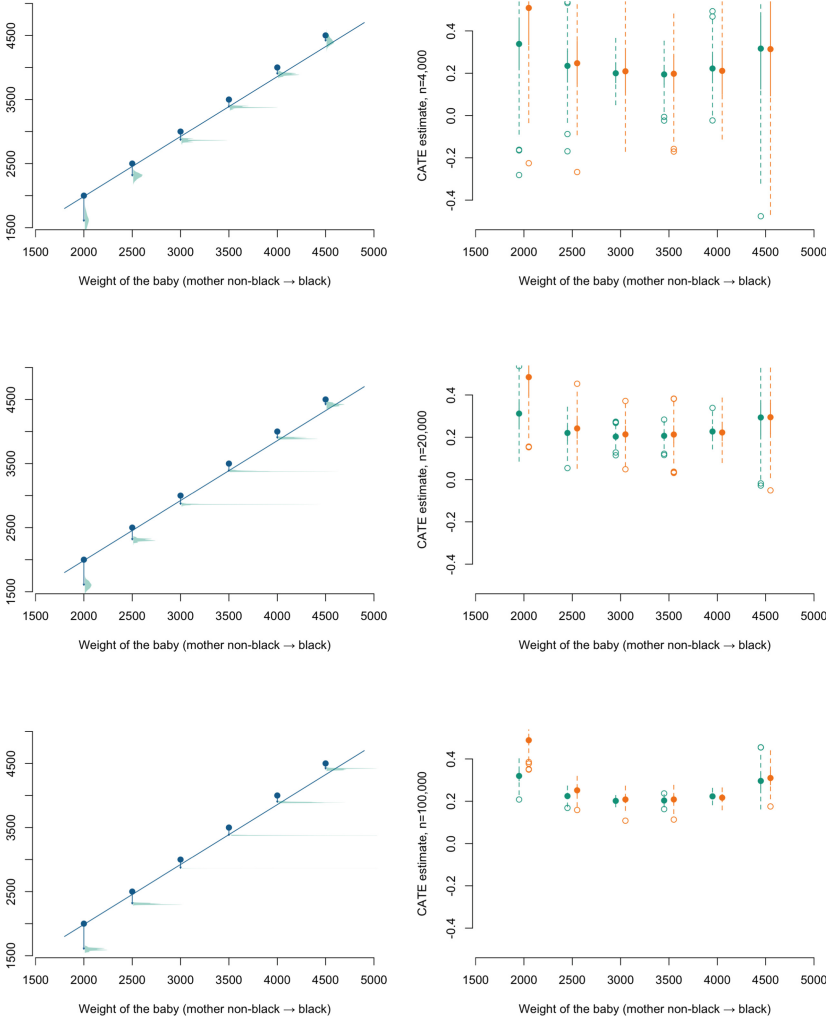
**Fig. 21.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using  $k$ -nearest neighbors, when  $Y = \mathbf{1}$ (non-natural delivery), and  $X$  is the weight gain of the mother, when  $T$  indicates whether the mother is a smoker or not. On the right, evolution of  $x \mapsto \text{SCATE}_{\mathcal{N}}[Y|X = x]$  with and without transport.



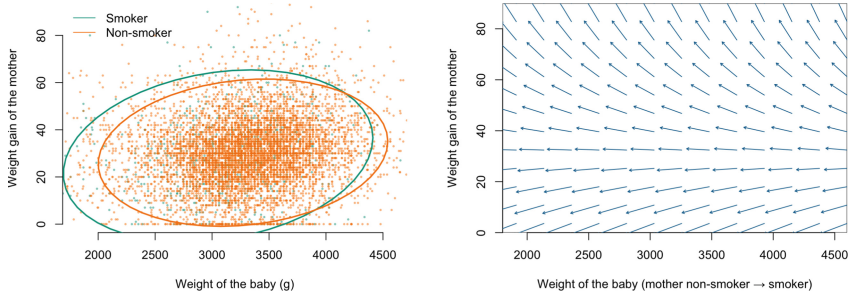
**Fig. 22.** On the left, evolution of  $x \mapsto \mathbb{E}[Y|X_{T \leftarrow t} = x, T = t]$ , estimated using a kernel based local average, when  $Y = \mathbf{1}$  (non-natural delivery), and  $X$  is the weight of the newborn infant, respectively when  $T$  indicates whether the mother is a smoker or not (on top), whether the mother is Black or not in the middle, and whether the newborn infant is a boy at the bottom. On the right, evolution of  $x \mapsto \text{SCATE}_{\mathcal{N}}[Y|X = x]$  with an without transport, based on a Gaussian transport.



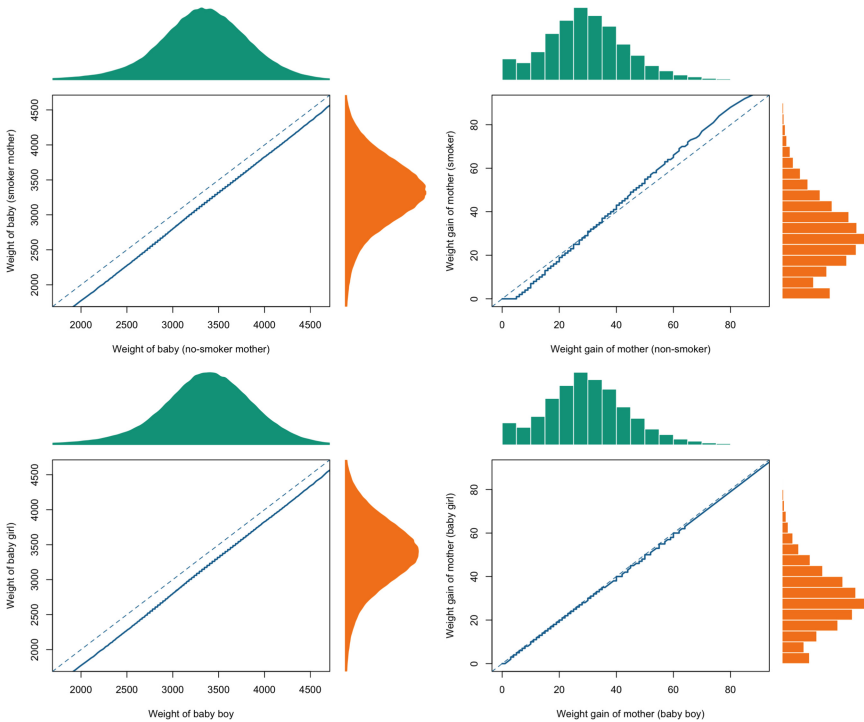
**Fig. 23.** On the left, distribution of  $\widehat{\mathcal{T}}(x)$ , when  $x$  is the weight of a newborn infant, when  $n$  goes to 4,000 (on top) to 20,000 (in the middle) and 100,000 (at the bottom), and when  $T$  indicates whether the mother is a smoker or not. On the right, boxplots of the estimation of  $\text{SCATE}(x)$ , with two GAM models, when  $x \in \{2000, 2500, \dots, 4000, 4500\}$ .



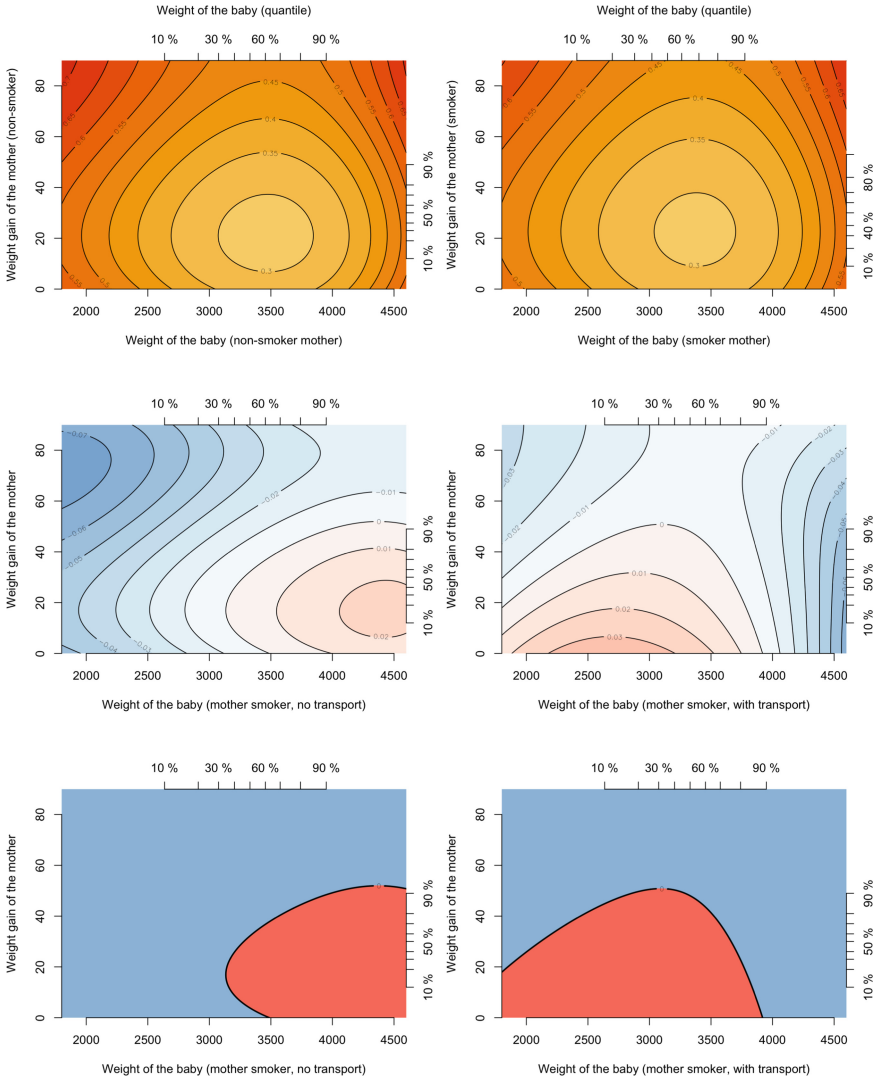
**Fig. 24.** On the left, distribution of  $\widehat{\mathcal{F}}(x)$ , when  $x$  is the weight of a newborn infant, when  $n$  goes to 4,000 (on top) to 20,000 (in the middle) and 100,000 (at the bottom), and when  $T$  indicates whether the mother is Afro-American or not. On the right, boxplots of the estimation of  $\text{SCATE}(x)$ , with two GAM models, when  $x \in \{2000, 2500, \dots, 4000, 4500\}$ .



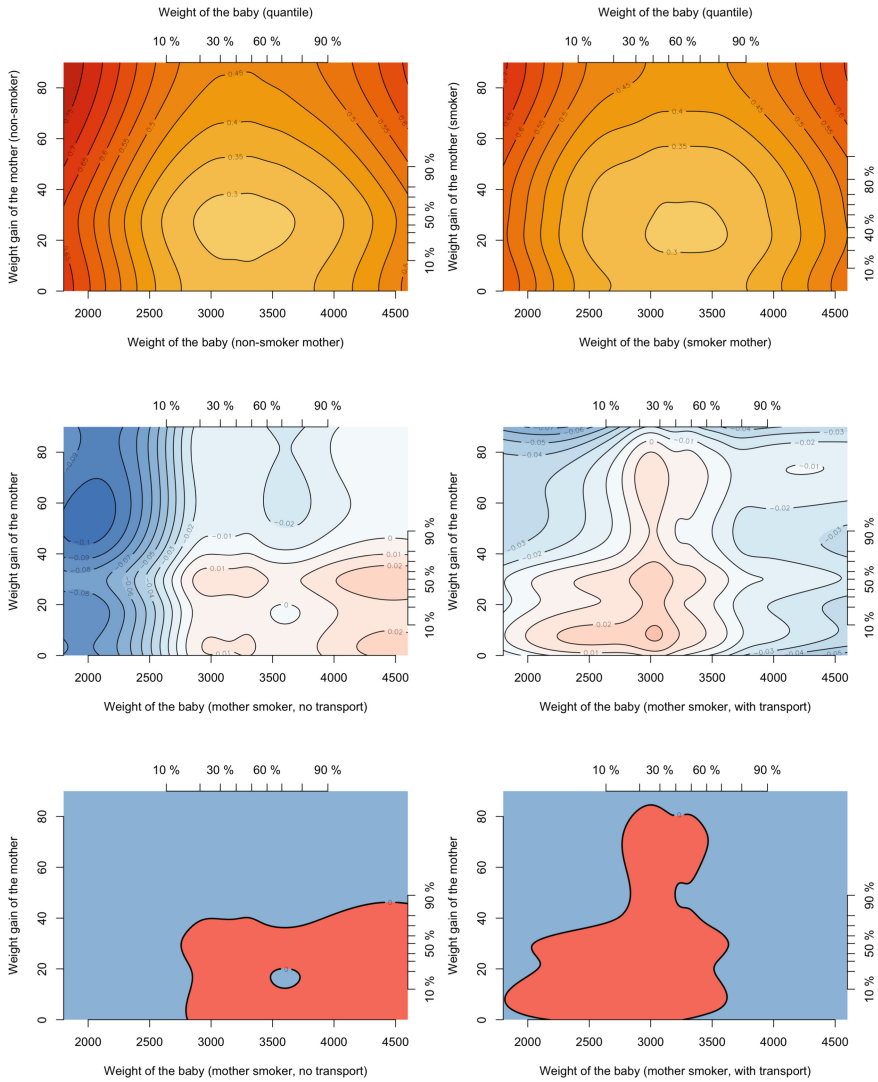
**Fig. 25.** Joint distributions of  $\mathbf{X}$  (weight of the newborn infant and weight gain of the mother), conditional on the treatment  $T$ , when  $T$  indicates whether the mother is a smoker or not on the left. On the right, vector field associated with optimal Gaussian transport, in dimension two (weight of the newborn infant and weight gain of the mother), when the treatment  $T$  indicates whether the mother is a smoker or not. Some numerical values are given in Table 7. On the right, the origin of the arrow is  $\mathbf{x}$  in the control group (non-smoker) and the arrowhead is  $\mathcal{T}_{\mathcal{N}}(\mathbf{x}$  in the treated group (smoker)).



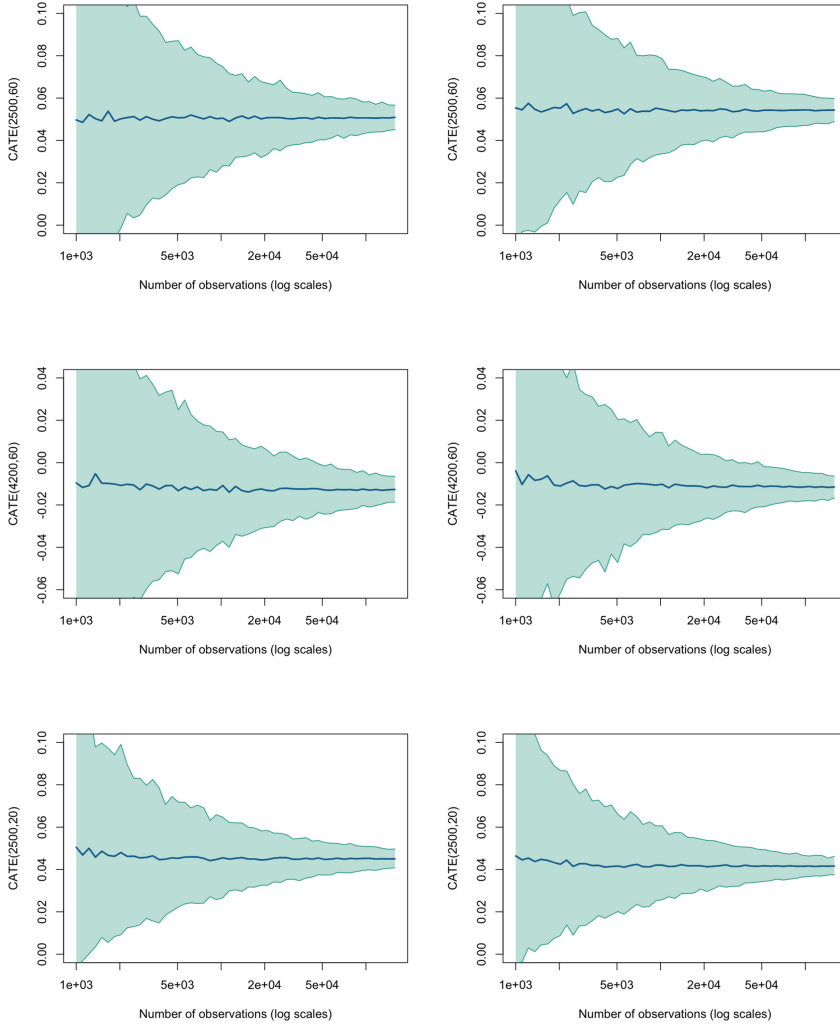
**Fig. 26.** Optimal transport (quantile based) when  $X$  is the weight of the newborn infant on the left, and the weight gain of the mother on the right, when  $T$  indicates whether the mother is a smoker on top, and whether the newborn infant is a boy at the bottom.



**Fig. 27.** On top, contours of  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0]$  and  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1]$  when  $T$  indicates whether a mother is a smoker or not, estimated with logistic GAM models (cubic splines). In the middle contours of the *ceteris paribus*  $\mathbf{x} \mapsto \text{CATE}[\mathbf{x}]$  without any transport on the left, and  $\mathbf{x} \mapsto \text{SCATE}[\mathbf{x}]$  *mutatis mutandis* on the right. At the bottom, positive/negative distinction for the conditional average treatment effect.

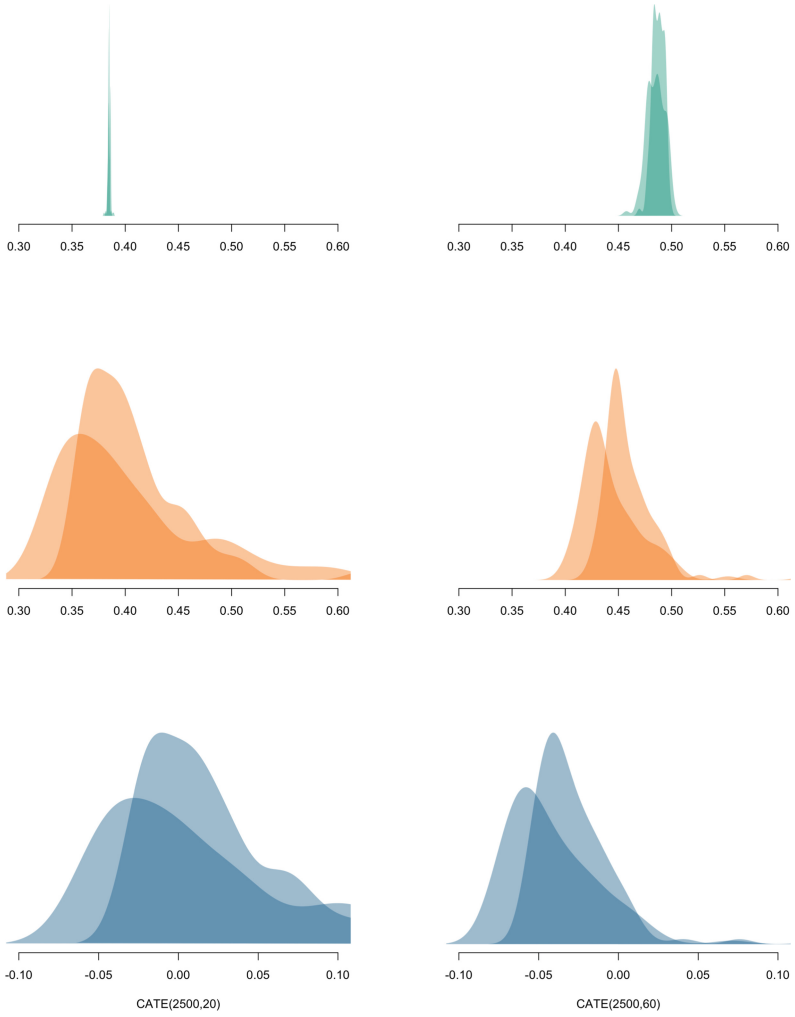


**Fig. 28.** On top, contours of  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0]$  and  $\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1]$  when  $T$  indicates whether a mother is a smoker or not, estimated with logistic GAM models (cubic splines, with more knots). In the middle contours of the *ceteris paribus*  $\mathbf{x} \mapsto \text{CATE}[\mathbf{x}]$  without any transport on the left, and  $\mathbf{x} \mapsto \text{SCATE}[\mathbf{x}]$  *mutatis mutandis* on the right. At the bottom, positive/negative effect for the conditional average treatment effect.



**Fig. 29.** On the left, estimation of  $\mathbf{x} \mapsto \text{SCATE}_{\mathcal{N}}[Y\mathbf{x}]$ , estimated using a logistic GAM model with Gaussian transport estimated on  $n$  observations (with  $n$  increasing from 1,000 to 150,000), when  $Y = \mathbf{1}$  (non-natural delivery), and  $\mathbf{X}$  is the weight of the newborn infant and the weight gain of the mother, respectively when  $T$  indicates whether the mother is a smoker or not (on the left), whether the mother is Black or not (on the right). On top  $\mathbf{x} = (2500, 60)$ , in the middle  $\mathbf{x} = (4200, 60)$  and at the bottom,  $\mathbf{x} = (2500, 20)$ .





**Fig. 30.** Distribution of  $\hat{m}_0(\mathbf{x})$  on top,  $\hat{m}_1(\widehat{\mathcal{T}}(\mathbf{x}))$  in the middle, and  $\text{SCATE}(\widehat{\mathcal{T}}(\mathbf{x})) = \hat{m}_1(\widehat{\mathcal{T}}(\mathbf{x})) - \hat{m}_0(\mathbf{x})$  at the bottom, estimated on bootstrapped samples of size  $n = 20,000$ . On the left,  $\mathbf{x} = (2500, 20)$  and on the right  $\mathbf{x} = (2500, 60)$ . The two densities are based on the fact that two GAM models are considered.

## References

- Abrevaya, J., Hsu, Y.-C., Lieli, R.P.: Estimating conditional average treatment effects. *J. Bus. Econ. Stat.* **33**(4), 485–505 (2015)
- Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. *Ann. Stat.* **47**(2), 1148–1178 (2019)
- Athey, S., Wager, S.: Estimating treatment effects with causal forests: an application. *Observational Stud.* **5**(2), 37–51 (2019)
- Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **44**(4), 375–417 (1991)
- Brualdi, R.A.: *Combinatorial Matrix Classes*, vol. 13. Cambridge University Press, Cambridge (2006)
- Charpentier, A.: Quantifying fairness and discrimination in predictive models. In: Kreinovich, V., SriboonchiNa, S., Yamaka, W. (eds.) *Machine Learning for Econometrics and Related Topics*. Springer Verlag (2023)
- Chisholm, R.M.: The contrary-to-fact conditional. *Mind* **55**(220), 289–307 (1946)
- Cunningham, S.: *Causal Inference*. Yale University Press, London (2021)
- Davis, J., Heller, S.B.: Using causal forests to predict treatment heterogeneity: an application to summer jobs. *Am. Econ. Rev.* **107**(5), 546–50 (2017)
- Dehejia, R.H., Wahba, S.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**(448), 1053–1062 (1999)
- Fan, Q., Hsu, Y.-C., Lieli, R.P., Zhang, Y.: Estimation of conditional average treatment effects with high-dimensional data. *J. Bus. Econ. Stat.* **40**(1), 313–327 (2022)
- Galichon, A.: *Optimal Transport Methods in Economics*. Princeton University Press, Princeton (2016)
- Goodman, N.: The problem of counterfactual conditionals. *J. Philos.* **44**(5), 113–128 (1947)
- Hahn, J.: On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331 (1998)
- Heckman, J.J., Ichimura, H., Todd, P.: Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* **65**(2), 261–294 (1998)
- Hernán, M.A., Robins, J.M.: *Causal inference*. CRC Press (2010)
- Hernández-Díaz, S., Schisterman, E.F., Hernán, M.A.: The birth weight “paradox” uncovered? *Am. J. Epidemiol.* **164**(11), 1115–1120 (2006)
- Higham, N.J.: *Functions of matrices: theory and computation*. In: SIAM (2008)
- Hitsch, G.J., Misra, S.: Heterogeneous treatment effects and optimal targeting policy evaluation (2018). Available at SSRN 3111957
- Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15**(3), 199–236 (2007)
- Hsu, Y.-C., Lai, T.-C., Lieli, R.P.: Counterfactual treatment effects: estimation and inference. *J. Bus. Econ. Stat.* **40**(1), 240–255 (2022)
- Imai, K.: *Quantitative Social Science: An Introduction*. Princeton University Press, Princeton (2018)
- Imbens, G.W., Rubin, D.B.: *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge (2015)
- Kantorovich, L.V.: On the translocation of masses. In: *Doklady Akademii Nauk USSR*, vol. 37, pp. 199–201 (1942)
- Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Nat. Acad. Sci.* **116**(10), 4156–4165 (2019)
- Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* **113**(521), 390–400 (2018)

- Mach, E.: The science of mechanics: a critical and historical exposition of its principles. Open court publishing Company (1893)
- McCann, R.J.: Exact solutions to the transportation problem on the line. *Proc. Royal Soc. London Ser. A Math. Phys. Eng. Sci.* **455**(1984), 1341–1380 (1999)
- Monge, G.: *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences de Paris (1781)
- Morgan, S.L., Winship, C.: *Counterfactuals and Causal Inference*. Cambridge University Press, Cambridge (2015)
- Pearl, J., Mackenzie, D.: *The Book of Why: The New Science of Cause and Effect*. Basic Books (2018)
- Powers, S., et al.: Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat. Med.* **37**(11), 1767–1787 (2018)
- Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983)
- Rubin, D.B.: Matching to remove bias in observational studies. *Biometrics* **29**, 159–183 (1973)
- Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**(5), 688 (1974)
- Santambrogio, F.: *Optimal transport for applied mathematicians*. Birkäuser, NY **55**(58–63), 94 (2015)
- Stuart, E.A.: Matching methods for causal inference: a review and a look forward. *Stat. Sci.* **25**(1), 1 (2010)
- Villani, C.: *Topics in Optimal Transportation*, vol. 58. American Mathematical Society, Providence (2003)
- Villani, C.: *Optimal Transport: Old and New*, vol. 338. Springer, Heidelberg (2009)
- Vowels, M.J., Camgoz, N.C., Bowden, R.: D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Comput. Surv. (CSUR)* **55**, 1–36 (2021)
- Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**(523), 1228–1242 (2018)
- Wilcox, A.J.: Birth weight and perinatal mortality: the effect of maternal smoking. *Am. J. Epidemiol.* **137**(10), 1098–1104 (1993)
- Wilcox, A.J.: On the importance-and the unimportance-of birthweight. *Int. J. Epidemiol.* **30**(6), 1233–1241 (2001)
- Marc, H., Eustasio, D.B., Juan, C.-A., Carlos, M.: Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Ann. Stat.* **49**(2), 1139–1165 (2021). <https://doi.org/10.1214/20-aos1996>



# Three Applications of Measure Transportation in Statistical Inference

Marc Hallin<sup>(✉)</sup>

ECARES and Department of Mathematics, Université libre de Bruxelles,  
Bruxelles, Belgium  
mhallin@ulb.ac.be

**Abstract.** Measure transportation recently helped solve some long-standing open problems in statistical inference by providing satisfactory definitions of ranks and quantiles beyond the classical univariate case. We describe three examples of this fact: Wilcoxon-type distribution-free rank-based tests for the multivariate two-sample problem with unspecified error densities, nonparametric multiple-output quantile regression, and Cramér-von Mises Goodness-of-Fit tests for directional data.

**Keywords:** Center-outward ranks and signs · Center-outward quantiles · Multivariate ranks and signs · Rank-based MANOVA · Multiple-output quantile regression · Directional Goodness-of-Fit

## 1 Introduction

Measure transportation in the past decades has emerged as one of the most active and vigorous subjects of pure and applied mathematics, with significant impact, basically, in all fields of applied mathematics, ranging from fluid dynamics to meteorology, economics, engineering design, operations research, and machine learning. Statistics was somewhat slower to join that tendency but recently quite successfully used measure transportation methods to solve several fundamental inference problems: see Galichon (2016), Panaretos and Zemel (2019) and Hallin (2022) for recent surveys.

A major success of measure transportation in statistical inference is the definition of distribution and quantile functions (and their empirical counterparts, ranks and empirical quantiles) beyond the traditional univariate case. In  $\mathbb{R}^d$  with  $d \geq 2$ , these definitions are based on gradients of convex functions pushing forward a distribution  $P$  under study either to the Lebesgue uniform over the unit cube  $[0, 1]^d$  or to the spherical uniform over the unit ball  $\mathbb{S}_d$ : see Chernozhukov et al. (2017), Hallin (2017), Hallin et al. (2021). In this paper, we are privileging transports to the unit ball, which, thanks to spherical symmetry, characterize attractive concepts of quantile regions and contours.

This measure transportation approach allows for the definition of the so-called *center-outward distribution* and *quantile functions*, *center-outward ranks*

and *signs*, and *center-outward empirical quantiles* for distributions over spaces for which no satisfactory forms of these concepts were available, such as  $\mathbb{R}^d$  (with  $d \geq 2$ ) or the unit sphere  $\mathcal{S}_{d-1}$  in  $\mathbb{R}^d$ . These concepts in turn provide solutions to a variety of problems: efficient rank-based distribution-free tests in  $\mathbb{R}^d$  and  $\mathcal{S}_{d-1}$ ; still in  $\mathbb{R}^d$  and  $\mathcal{S}_{d-1}$  ( $d \geq 2$ ), construction of (conditional) quantile contours and regions; multiple-output quantile regression; nonparametric inference for directional data; etc. We illustrate that fact here with three examples: Wilcoxon tests for the multivariate two-sample problem with unspecified absolutely continuous error densities, nonparametric multiple-output quantile regression, and Cramér-von Mises Goodness-of-Fit tests on the (hyper)sphere.

## 2 Measure-Transportation-Based Center-Outward Distribution and Quantile Functions, Ranks, and Signs in $\mathbb{R}^d$

Traditional (univariate) ranks are based on a monotone increasing mapping (the empirical distribution function  $F^{(n)}$ ) of a sample  $Z_1, \dots, Z_n$  to a regular grid of the form<sup>1</sup>  $\mathfrak{G}_1^{(n)} := \{\frac{1}{n+1}, \dots, \frac{n}{n+1}\}$ ; the discrete uniform distribution over  $\mathfrak{G}_{1;\pm}^{(n)}$  converges weakly as  $n \rightarrow \infty$ , to the uniform over the unit interval  $[0, 1]$ . The rank of  $Z_i$  then is  $R_i^{(n)} := (n + 1)F^{(n)}(Z_i)$  and corresponds to the ordering of the observations from smallest to largest—the complete ordering, from  $-\infty$  to  $\infty$ , of the real line  $\mathbb{R}$ .

Such ordering cannot be expected to extend to dimension two and higher where, instead of  $-\infty$  and  $\infty$ , each direction (each point on the unit hypersphere  $\mathcal{S}_{d-1}$ ) characterizes a point at infinity, suggesting a center-outward (partial) ordering rather than a “left-to-right” complete one. Therefore, still in the univariate case, let us replace the traditional empirical distribution function  $F^{(n)}$  with the so-called center-outward empirical distribution function  $F_{\pm}^{(n)}$  mapping the sample values to a regular grid of the open interval  $\mathbb{S}_1 := (-1, 1)$ —which is the unit ball in  $\mathbb{R}$ —of the form

$$\mathfrak{G}_{1;\pm}^{(n)} := \left\{ \frac{\pm 1}{\lfloor n/2 \rfloor + 1}, \dots, \frac{\pm \lfloor n/2 \rfloor}{\lfloor n/2 \rfloor + 1} \right\} \quad \text{along with the origin if } n \text{ is odd;}$$

note that the discrete uniform distribution over  $\mathfrak{G}_{1;\pm}^{(n)}$  converges weakly, as  $n \rightarrow \infty$ , to the uniform  $U_1$  over the unit ball  $[-1, 1]$ . The *center-outward rank* of  $Z_i$  then is defined as  $R_{i;\pm}^{(n)} := \left| (\lfloor n/2 \rfloor + 1)F_{\pm}^{(n)}(Z_i) \right|$ , and the unit vector  $S_{i;\pm}^{(n)} := F_{\pm}^{(n)}(Z_i) / \left| F_{\pm}^{(n)}(Z_i) \right|$  (taking values  $\pm 1$  and 0 at the origin) is a *center-outward sign*.

Being monotone increasing,  $F^{(n)}$  (respectively,  $F_{\pm}^{(n)}$ ) is minimizing  $\sum_{i=1}^n [T(Z_i) - Z_i]^2$  over all pairings  $T$  between  $\{Z_1, \dots, Z_n\}$  and  $\mathfrak{G}_1^{(n)}$  (respec-

<sup>1</sup> As usual in rank-based inference, the denominator, in the definition of  $F^{(n)}$ , is chosen as  $(n + 1)$  rather than  $n$ : the range of  $F^{(n)}$  then is in the *open* unit interval  $(0, 1)$ .

tively,  $\mathfrak{G}_{1;\pm}^{(n)}$  and thus is an optimal transport pushing the empirical distribution of the sample forward to the empirical distribution of the grid. Clearly, the traditional ranks  $\{R_1^{(n)}, \dots, R_n^{(n)}\}$  and the center-outward ranks and signs  $\{R_{1;\pm}^{(n)}, S_{1;\pm}^{(n)}, \dots, R_{n;\pm}^{(n)}, S_{N;\pm}^{(n)}\}$  are functions of each other, hence carry the same information and generate the same sigma-fields; all traditional univariate rank statistics thus can be rewritten in terms of the center-outward ranks and signs.

Turning to  $d \geq 2$ , the above definitions naturally extend, now based on a regular grid  $\mathfrak{G}_{d;\pm}^{(n)}$  of the  $d$ -dimensional unit ball  $\mathbb{S}_d$  with the property that the uniform discrete distribution over  $\mathfrak{G}_{d;\pm}^{(n)}$  converges weakly, as  $n \rightarrow \infty$ , to the spherical uniform over  $\mathbb{S}_d$ . The empirical center-outward distribution function  $\mathbf{F}_{\pm}^{(n)}$  of a sample  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$  then is obtained as the minimizer, over all possible bijections  $\mathbf{T}$  between  $\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$  and the grid  $\mathfrak{G}_{d;\pm}^{(n)}$ , of  $\sum_{i=1}^n \|\mathbf{T}(\mathbf{Z}_i) - \mathbf{Z}_i\|^2$ , that is, as the optimal transport pushing the empirical distribution of the sample forward to the empirical distribution of the grid.

If center-outward ranks (with integer values  $1, 2, \dots$ ), center-outward signs (with modulus one), and empirical center-outward quantile contours (not consisting of one single point) are to be defined, a special type of grid can be adopted as follows. Factorizing  $n$  into  $n_R n_S + n_0$  with  $n_0 < \min(n_R, n_S)$ , consider a regular array of  $n_S$  directions ( $n_S$  points on the unit sphere  $\mathcal{S}_{d-1}$ : for instance,  $n_S$  i.i.d. points uniformly distributed over  $\mathcal{S}_{d-1}$ ). Then take the intersections between the corresponding rays and the nested hyperspheres with radii  $\frac{1}{n_R+1}, \dots, \frac{n_R}{n_R+1}$ : the grid  $\mathfrak{G}_{d;\pm}^{(n)}$  consists of the resulting  $n_R n_S$  points, along with  $n_0$  copies of the origin: see Fig. 1 for an example in dimension  $d = 2$ . The *center-outward rank* and *center-outward sign* of  $\mathbf{Z}_i$  then are defined as  $R_{i;\pm}^{(n)} := (n_R + 1) \left\| \mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_i) \right\|$  and  $\mathbf{S}_{i;\pm}^{(n)} := \mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_i) / \left\| \mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_i) \right\|$ , respectively, with the convention that  $\mathbf{S}_{i;\pm}^{(n)} = \mathbf{0}$  for  $\mathbf{Z}_i^{(n)} = \mathbf{0}$ ; the empirical *center-outward quantile contour* of  $\mathbf{Z}_i$  naturally consists of all observations  $\mathbf{Z}_j$  such that  $R_{j;\pm}^{(n)} = R_{i;\pm}^{(n)}$ .

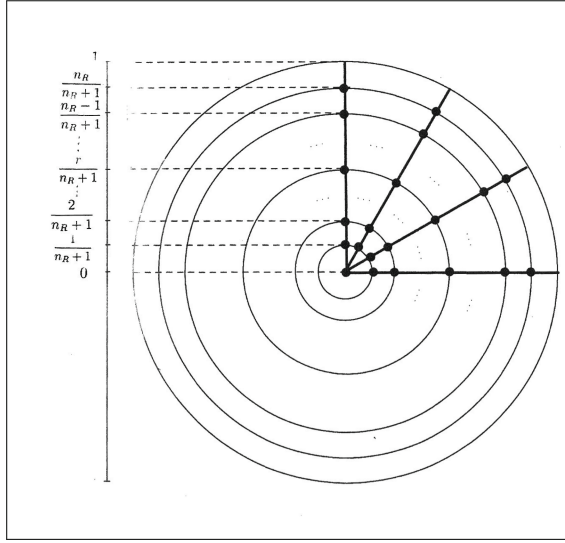
For simplicity, let us assume throughout that the random vectors  $\mathbf{Z}_i$  are i.i.d. with Lebesgue-absolutely continuous distribution  $P$  and density  $f$  such that, for any closed ball of the form  $c\bar{\mathbb{S}}_d (c_i \mathbf{0})$  there exist constants  $B_c^-$  and  $B_c^+$  such that<sup>2</sup>

$$0 < B_c^- \leq f(\mathbf{z}) \leq B_c^+ < \infty \quad \text{for all } \mathbf{z} \in c\bar{\mathbb{S}}_d ; \tag{1}$$

denote by  $\mathcal{P}_d^*$  the collection of distributions satisfying (1). The population version  $\mathbf{F}_{\pm}$  of the center-outward distribution function of  $P \in \mathcal{P}_d^*$  is defined as the (a.s. unique) gradient of a convex function<sup>3</sup> pushing  $P$  forward to  $U_d$  and has been shown (Figalli 2018) to be a homeomorphism between  $\mathbb{S}_d \setminus \{\mathbf{0}\}$  and  $\mathbb{R}^d \setminus \mathbf{F}_{\pm}(\mathbf{0})$ . Therefore,  $\mathbf{F}_{\pm}$  admits a continuous inverse  $\mathbf{Q}_{\pm}$  (possibly set-

<sup>2</sup> That assumption can be relaxed: see del Barrio et al. (2020).

<sup>3</sup> Existence and a.s. uniqueness of  $\mathbf{F}_{\pm}$  are guaranteed by the results of McCann (1995). Whenever  $P$  has finite second-order moments,  $\mathbf{F}_{\pm}$  moreover is the optimal (for quadratic costs) transport pushing  $P$  forward to  $U_d$ .



**Fig. 1.** A typical grid  $\mathfrak{G}_{d;\pm}^{(n)}$  in dimension  $d = 2$ .

valued at  $\mathbf{0}$ ) which pushes  $U_d$  forward to  $P$  and qualifies as the *center-outward quantile function* of  $P$ .

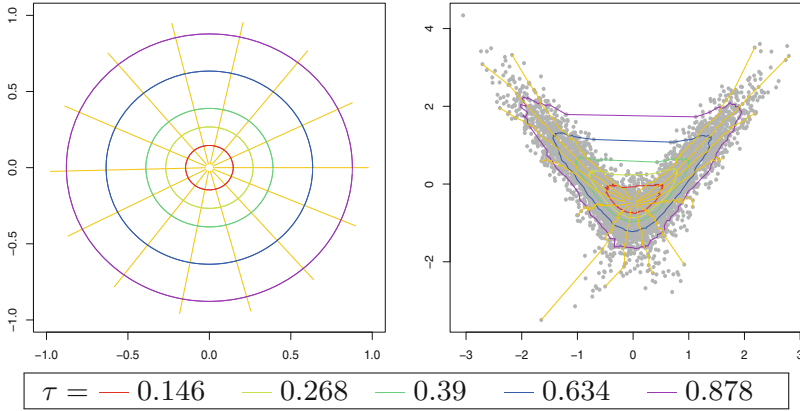
It is well known that optimal transports, in general, cannot be expressed under analytical form. However, it can be shown that, if  $\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_n^{(n)}$  are i.i.d. with distribution  $P$  satisfying (1) and provided that the empirical distribution over  $\mathfrak{G}_{d;\pm}^{(n)}$  converges weakly to the uniform  $U_d$ ,  $\mathbf{F}_{\pm}^{(n)}$  and  $\mathbf{F}_{\pm}$  are related by the Glivenko-Cantelli result

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} \left\| \mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_i^{(n)}) - \mathbf{F}_{\pm}(\mathbf{Z}_i^{(n)}) \right\| = 0 \quad \text{a.s.}$$

This allows us to construct arbitrarily precise numerical evaluations of  $\mathbf{F}_{\pm}$  (equivalently, of  $\mathbf{Q}_{\pm} := \mathbf{F}_{\pm}^{-1}$ ). Fig. 2 illustrates this fact in dimension  $d = 2$ .<sup>4</sup>

We refer to Hallin et al. (2021) for details and complements.

<sup>4</sup> Figure prepared by Gilles Mordant.



**Fig. 2.** Left panel: the two-dimensional unit ball, with the nested spheres with radii  $\tau = 0.146, 0.268, 0.390, 0.634,$  and  $0.878$  (the balls  $\tau\bar{\mathbb{S}}_d$  with radius  $\tau$  centered at the origin). Right-hand panel: a numerical approximation, based on a sample of points generated from a banana-shaped mixture  $P$  of three normals, of the corresponding quantile regions  $\mathbf{Q}_\pm(\tau\bar{\mathbb{S}}_d)$  (same  $\tau$  values) of  $P$ . The  $P$ -probability content of  $\mathbf{Q}_\pm(\tau\bar{\mathbb{S}}_d)$  is the  $U_d$ -probability content of the ball  $\tau\bar{\mathbb{S}}_d$ , which is  $\tau$ , irrespective of  $P$ .

### 3 A Wilcoxon Test for the Multivariate Two-Sample Problem

Wilcoxon’s two-sample test is perhaps the most classical example of a univariate rank test. A substitute for Student’s two-sample test, it goes back to Wilcoxon’s foundational paper (Wilcoxon 1945), which is generally considered as the starting point of rank-based inference.

Denoting by  $(Z_1^{(n)}, \dots, Z_{n_1}^{(n)}, Z_{n_1+1}^{(n)}, \dots, Z_n^{(n)})$  an  $n$ -tuple of univariate independent observations with

$$Z_1^{(n)} - \delta, \dots, Z_{n_1}^{(n)} - \delta, Z_{n_1+1}^{(n)}, \dots, Z_n^{(n)} \quad \text{i.i.d. } P$$

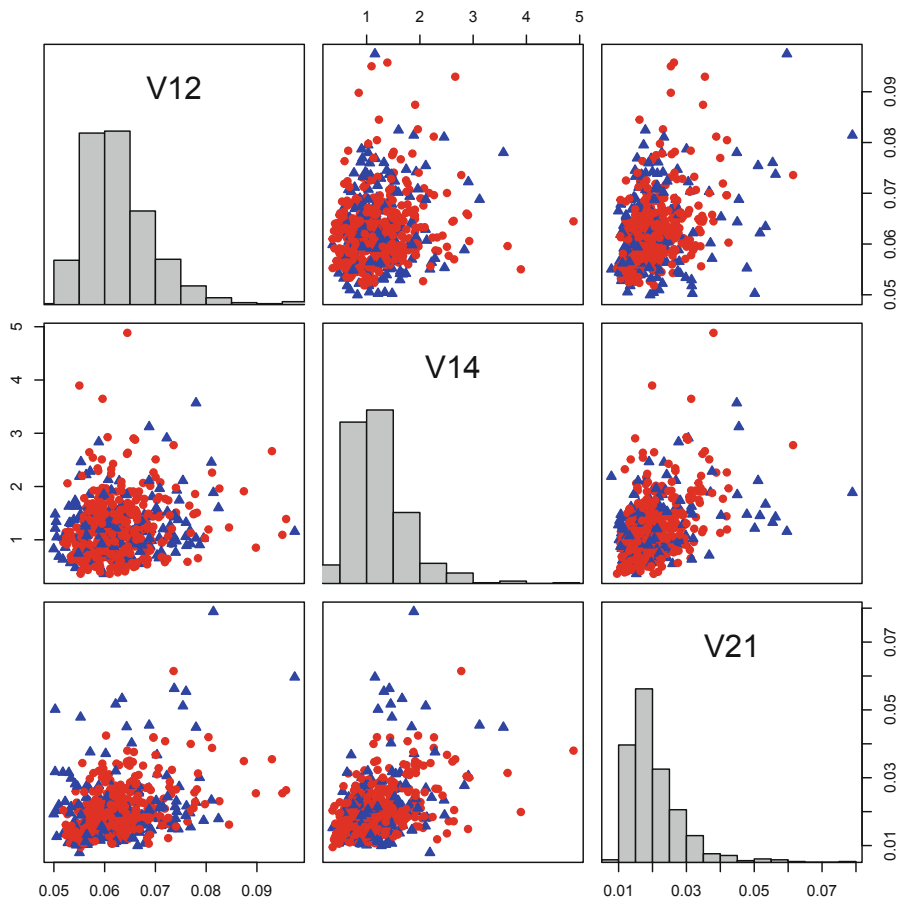
where  $\delta \in \mathbb{R}$  and  $P$  is an absolutely continuous but otherwise unspecified distribution, the typical hypothesis to be tested in the two-sample problem is  $\mathcal{H}_0^{(n)} : \delta = 0$  (no treatment effect) against the one-sided alternative  $\mathcal{H}_{1,+}^{(n)} : \delta > 0$  (strictly positive treatment effect) or the two-sided one  $\mathcal{H}_1^{(n)} : \delta \neq 0$ . The one-sided Wilcoxon test is based on the Wilcoxon statistic

$$\zeta_w^{(n)} := \left( \frac{n_1 n_2 (n + 1)}{12} \right)^{-1/2} \left( \sum_{i=1}^{n_1} R_i^{(n)} - \frac{n_1 (n + 1)}{2} \right).$$

where  $n_2 := n - n_1$ . The null distribution of  $\zeta_w^{(n)}$  is tabulated (see, e.g., Lehmann (1975)); asymptotically, as  $n_1$  and  $n_2$  both tend to infinity,  $\zeta_w^{(n)}$  is



asymptotically standard normal, and  $\mathcal{H}_0^{(n)}$  is rejected in favor of  $\mathcal{H}_{1;+}^{(n)}$  at asymptotic level<sup>5</sup>  $\alpha$  whenever  $\mathcal{Z}_w^{(n)}$  exceeds the standard normal quantile of order  $(1-\alpha)$ . The quadratic test statistic  $Q_w^{(n)} := \left(\mathcal{Z}_w^{(n)}\right)^2$ , with asymptotically chi-square (one degree of freedom) null distribution, similarly can be used against the two-sided alternative  $\mathcal{H}_1^{(n)}$ .



**Fig. 3.** Wisconsin Diagnostic Breast Cancer (WDBC) data: bivariate scatterplots and univariate histograms for mean fractal dimension (V12), standard error of texture (V14), and standard error of symmetry (V21) in 212 malignant (triangles) and 357 benign patients (circles).

<sup>5</sup> Thanks to distribution-freeness under  $\mathcal{H}_0^{(n)}$ , the size of the test is uniform with respect to the underlying P. So is thus the convergence to  $\alpha$  of that size.

In dimension  $d > 1$ , for a sample  $(\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_{n_1}^{(n)}, \mathbf{Z}_{n_1+1}^{(n)}, \dots, \mathbf{Z}_n^{(n)})$  satisfying

$$\mathbf{Z}_1^{(n)} - \boldsymbol{\delta}, \dots, \mathbf{Z}_{n_1}^{(n)} - \boldsymbol{\delta}, \mathbf{Z}_{n_1+1}^{(n)}, \dots, \mathbf{Z}_n^{(n)} \quad \text{i.i.d. P}$$

with  $\boldsymbol{\delta} \in \mathbb{R}^d$  and  $P \in \mathcal{P}_*^{(n)}$  otherwise unspecified, the null hypothesis takes the form  $\mathcal{H}_0^{(n)} : \boldsymbol{\delta} = \mathbf{0}$  (no treatment effect), to be tested against the alternative  $\mathcal{H}_1^{(n)} : \boldsymbol{\delta} \neq \mathbf{0}$ . With the notation of Sect. 2, a Wilcoxon center-outward test statistic can be defined as

$$\mathbf{Q}_{\pm;w}^{(n)} := \frac{3nd}{n_1 n_2 (n_R + 1)^2} \left\| \sum_{i=1}^{n_1} R_{i;\pm}^{(n)} \mathbf{S}_{i;\pm}^{(n)} \right\|^2$$

which, for  $d = 1$ , is asymptotically equivalent to  $Q_w^{(n)}$  under  $\mathcal{H}_0^{(n)}$  and contiguous alternatives. We refer to Hallin et al. (2022a) and Hallin and Mordant (2023) for details, a simulation study of finite-sample performance, and empirical evidence of the power of this test, particularly under skewed multivariate distributions.

The following real-life example illustrates the importance of center-outward rank tests. The Wisconsin Diagnostic Breast Cancer data (WDBC; dataset available at Machine Learning Repository (Dua and Graff (2017))) contains records of  $n = 569$  patients from two groups—benign or malignant tumor diagnosis. For each patient, several features were recorded from the digitized image of a fine needle aspirate of the breast mass, resulting in  $d = 30$  variables, among which mean fractal dimension (V12), standard error of texture (V14), and standard error of symmetry (V21). The classical testing method for discriminating between the two groups is the Hotelling test which, with a  $p$ -value of 0.9899, is highly inconclusive. In sharp contrast, the center-outward Wilcoxon rank test yields a  $p$ -value of 0.0327. The difference is explained by the poor resistance to skewness (see Fig. 3)<sup>6</sup> of Hotelling—a feature that does not affect the distribution-freeness of Wilcoxon.

The relevance of the method in medical diagnosis aid is quite obvious.

## 4 Multiple-Output Quantile Regression

Quantile regression methods are perhaps the most powerful and most informative tool in the study of the dependence of a variable of interest  $Y$  on covariates  $\mathbf{X} = (X_1, \dots, X_m)$ . Since their introduction by Koenker and Bassett (1978), they have become part of daily statistical practice, with countless applications in all domains of scientific research, from economics and social sciences to astronomy, biostatistics, and medicine. Unlike classical regression, which narrowly focuses on conditional means  $E[Y|\mathbf{X}]$ , quantile regression indeed is dealing with the complete conditional distributions  $P_{Y|\mathbf{X}=\mathbf{x}}$  of  $Y$  conditional on  $\mathbf{X} = \mathbf{x}$ . A number of quantile regression techniques, parametric, semiparametric, and nonparametric, have been developed for an extremely broad range

<sup>6</sup> Figure prepared by Šarka Hudecová.

of statistical models. We refer to Koenker (2005) for an introductory text and to Koenker et al. (2018) for a comprehensive survey.

Quantile regression thus is well understood and well developed in single-output models (univariate variable of interest  $Y$ ); results are much scarcer, however, in the ubiquitous multiple-output case ( $d$ -dimensional variable of interest  $\mathbf{Y}$ , with  $d > 1$ ), and the few existing results (e.g., based on conditional depth: see Hallin and Šiman (2018)) are less satisfactory<sup>7</sup>—the simple reason for this being the absence of a fully satisfactory concept of multivariate quantiles.

The measure-transportation-based concepts introduced in Sect. 2 allow for a new approach, which has been developed in del Barrio et al. (2022). The center-outward quantile function  $\mathbf{Q}_\pm$  of an absolutely continuous distribution  $P$  over  $\mathbb{R}^d$  defines nested closed connected quantile regions<sup>8</sup>  $\mathbb{C}_P(\tau) := \mathbf{Q}_\pm(\tau \bar{\mathbb{S}}_d)$  and continuous contours  $\mathbb{C}_P(\tau) := \mathbf{Q}_\pm(\tau \mathcal{S}_{d-1})$  indexed by  $\tau \in ([0, 1])$  such that, irrespective of the actual distribution  $P$ , the probability contents  $P[\mathbb{C}_P(\tau)]$  of  $\mathbb{C}_P(\tau)$  is  $\tau$  for all  $\tau \in ([0, 1])$ . Unlike the depth-based concepts (considered, e.g., in Hallin et al. (2010, 2015) or Kong and Mizera (2012)), thus, these measure-transportation-based quantiles do satisfy the essential property that the  $P$ -probability contents of the resulting quantile regions do not depend on  $P$ . Moreover, the corresponding quantile regions are closed, connected, and nested; they are not necessarily convex and are able to capture the “shape” of the underlying distribution.

Considering a pair of multidimensional random variables  $(\mathbf{X}, \mathbf{Y})$  with values in  $\mathbb{R}^m \times \mathbb{R}^d$  ( $\mathbf{Y}$  the variable of interest,  $\mathbf{X}$  the vector of covariates) and joint distribution<sup>9</sup>  $P$ , define the *center-outward quantile map*  $\mathbf{Q}_\pm(\cdot | \mathbf{x})$  of  $\mathbf{Y}$  conditional on  $\mathbf{X} = \mathbf{x}$  as the center-outward quantile function of the distribution of  $\mathbf{Y}$  conditional on  $\mathbf{X} = \mathbf{x}$ , namely,

$$\mathbf{u} \in \mathbb{S}_d \mapsto \mathbf{Q}_\pm(\mathbf{u} | \mathbf{x}) \in \mathbb{R}^d \tag{2}$$

( $\mathbb{S}_d$  the open unit ball in  $\mathbb{R}^d$ ). That mapping enjoys the essential property that, letting

$$\mathbb{C}_\pm(\tau | \mathbf{x}) := \mathbf{Q}_\pm(\tau \bar{\mathbb{S}}_d | \mathbf{x}) \quad \tau \in (0, 1), \quad \mathbf{x} \in \mathbb{R}^m, \tag{3}$$

we have

$$\mathbb{P}[\mathbf{Y} \in \mathbb{C}_\pm(\tau | \mathbf{x}) | \mathbf{X} = \mathbf{x}] = \tau \quad \text{for all } \mathbf{x} \in \mathbb{R}^m, \tau \in (0, 1), \text{ and } \mathbb{P}, \tag{4}$$

justifying the interpretation of  $\mathbf{x} \mapsto \mathbb{C}_\pm(\tau | \mathbf{x})$  as the value at  $\mathbf{x}$  of a *regression quantile region of order  $\tau$*  of  $\mathbf{Y}$  with respect to  $\mathbf{X}$ . For  $\tau = 0$ ,

$$\mathbb{C}_\pm(0 | \mathbf{x}) := \bigcap_{\tau \in (0, 1)} \mathbb{C}_\pm(\tau | \mathbf{x}) \tag{5}$$

<sup>7</sup> The main drawback of depth-based quantile regions is that their probability contents are not under control, and depend on the underlying distributions.

<sup>8</sup> We denote by  $\bar{\mathbb{S}}_d$  the closed unit ball in  $\mathbb{R}^d$ .

<sup>9</sup> For simplicity, we tacitly assume all distributions to be Lebesgue-absolutely continuous.

yields the value at  $\mathbf{X} = \mathbf{x}$  of the *center-outward regression median*  $\mathbf{x} \mapsto \mathbb{C}_{\pm}(0 | \mathbf{x})$  of  $\mathbf{Y}$  with respect to  $\mathbf{X}$ . The same conditional quantile map characterizes nested (no “quantile crossing” phenomenon) “*regression quantile tubes of order  $\tau$* ” (in  $\mathbb{R}^{m+d}$ )

$$\mathbb{T}_{\pm}(\tau) := \{(\mathbf{x}, \mathbf{Q}_{\pm}(\tau \bar{\mathbb{S}}_d | \mathbf{x})) \mid \mathbf{x} \in \mathbb{R}^m\}, \quad \tau \in (0, 1) \tag{6}$$

which are such that

$$\mathbb{P}[(\mathbf{X}, \mathbf{Y}) \in \mathbb{T}_{\pm}(\tau)] = \tau \quad \text{irrespective of } \mathbb{P}, \tau \in (0, 1). \tag{7}$$

For  $\tau = 0$ , define

$$\mathbb{T}_{\pm}(0) := \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{C}_{\pm}(0 | \mathbf{x})\} = \bigcap_{\tau \in (0,1)} \mathbb{T}_{\pm}(\tau)$$

(the *graph* of  $\mathbf{x} \mapsto \mathbb{C}_{\pm}(\tau | \mathbf{x})$ ); with a slight abuse of language, also call  $\mathbb{T}_{\pm}(0)$  the *regression median* of  $\mathbf{Y}$  with respect to  $\mathbf{X}$ . These concepts are developed in del Barrio et al. (2022), where consistent estimators are also proposed.

As usual in the context of measure transportation, no analytical form is possible for the conditional center-outward quantile mappings  $\mathbf{Q}_{\pm}(\cdot | \mathbf{x})$  and the resulting quantile regression contours, regions, and tubes. Arbitrarily precise numerical approximations can be constructed, though, from simulated samples with adequately large size  $n$ . Figure 5 provides an example of such an approximation for  $n = 55022$  i.i.d. simulations of  $(X, \mathbf{Y})$  where  $X$  is uniform over  $(-2, 2)$  and, conditional on  $X = x$ , the distribution of  $\mathbf{Y} = (Y_1, Y_2)$  is that of

$$(x, x^2) + (1 + 3 \sin^2(\pi x/2)) [(0, 1) + \boldsymbol{\xi}] \tag{8}$$

(note the quadratic-in- $x$  trend and periodic-in- $x$  heteroskedasticity) with  $\boldsymbol{\xi}$  a mixture of three bivariate normal distributions (as in Fig. 2), with density

$$\frac{3}{8} f_{\mathcal{N}((-3,0), \boldsymbol{\Sigma}_1)} + \frac{1}{4} f_{\mathcal{N}((0,0), \boldsymbol{\Sigma}_2)} + \frac{3}{8} f_{\mathcal{N}((3,0), \boldsymbol{\Sigma}_3)} \tag{9}$$

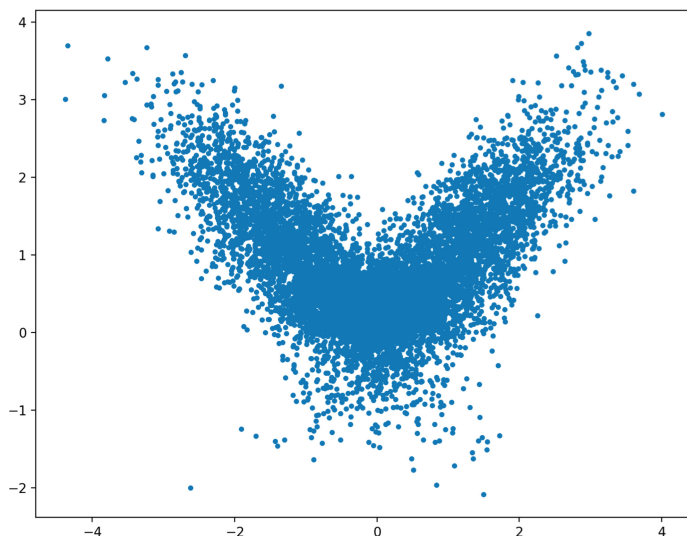
(writing  $f_{\mathcal{N}(((\mu_1, \mu_2)', \boldsymbol{\Sigma}))}$  for the density of the bivariate  $\mathcal{N}(((\mu_1, \mu_2)', \boldsymbol{\Sigma}))$  variable) and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma}_3 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}. \tag{10}$$

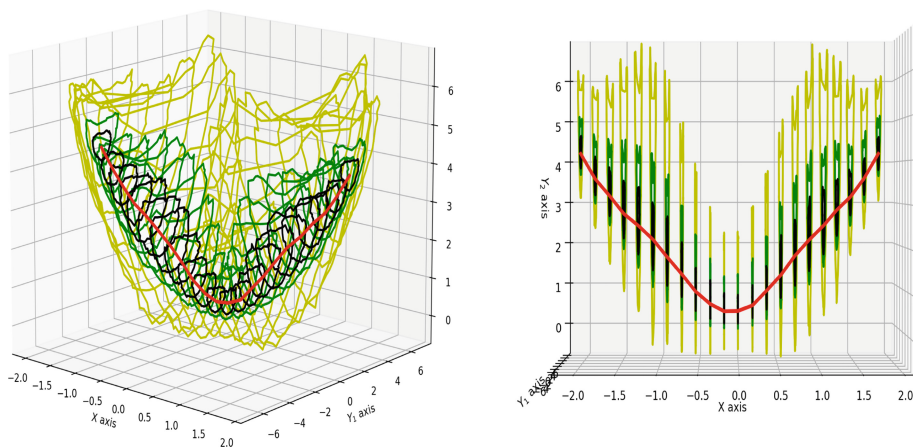
The simulated values of  $\boldsymbol{\xi}$  are shown in Fig. 4<sup>10</sup>, yielding a non-convex, banana-shaped distribution; the resulting numerical approximation of center-outward quantile regression contours, regions, and tubes (orders  $\tau = 0.2, 0.4$ , and  $0.8$ ), along with the center-outward regression median, are shown in Fig. 5.<sup>10</sup> We refer to del Barrio et al. (2022) for details.

---

<sup>10</sup> Figure prepared by Alberto González-Sanz.



**Fig. 4.** Simulated sample of  $n = 55022$  i.i.d. copies of  $\xi$  described in (9)–(10).



**Fig. 5.** Two perspectives on the numerical approximation of the center-outward regression median (in red) and center-outward quantile regression tubes (orders  $\tau = 0.2$  (black),  $0.4$  (green), and  $0.8$  (yellow)) of  $\mathbf{Y}$  with respect to  $X$  as described in equations (8)–(10).

## 5 Cramér-von Mises Goodness-of-Fit Tests for Directional Data

Goodness-of-Fit tests for univariate data typically are based on distances between their empirical distribution function  $F^{(n)}$  and some null distribution function  $F_0$ :  $\max_{1 \leq i \leq n} |F^{(n)}(Z_i^{(n)}) - F_0(Z_i^{(n)})|$  for the Kolmogorov-Smirnov

test,  $\sum_{i=1}^n \left| F^{(n)}(Z_i^{(n)}) - F_0(Z_i^{(n)}) \right|^2$  for the Cramér-von Mises test,  $W_2(F^{(n)}, F_0)$  for the Wasserstein-distance-based tests (see Hallin, Mordant, and Segers 2021), etc. These tests all reject  $\mathcal{H}_0 : P = P_0$  (where  $P$  denotes the actual distribution of the observations,  $P_0$  the distribution characterized by  $F_0$ ) for large values of these distances.

This is fine for real-valued observations  $Z_1^{(n)}, \dots, Z_n^{(n)}$ , with distribution function  $F$  and empirical distribution function  $F^{(n)}$ . What about directional observations  $\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_n^{(n)}$  taking values on the hypersphere  $\mathcal{S}_{d-1}$  in  $\mathbb{R}^d$ ? In dimension  $d = 3$  and higher,  $\mathcal{S}_{d-1}$  does not admit any canonical ordering and the concepts of distribution and quantile functions become problematic: what are  $F_0$ ,  $F$ ,  $F^{(n)}$  there?

Goodness-of-Fit problems for directional data nevertheless have a number of real-life applications, and have been considered extensively. This includes, among other developments,

- (i) Sobolev tests of uniformity on  $\mathcal{S}^{d-1}$  based on the eigenvectors associated with the non-zero eigenvalues of the Laplacian operator on  $\mathcal{S}^{d-1}$  (Jupp (2005, 2008), Jammalamadaka Rao et al. (2020)); projection-based tests of the Cramér-von Mises, Anderson-Darling, and Rothman type (García-Portugués et al. (2020, 2023));
- (ii) depth- and quantile-based tests based on angular simplicial depth and angular Tukey depth (Liu and Singh (1992), Rousseeuw and Struyf (2004) and Agostinelli and Romanazzi (2013)); tests based on the quantiles of sample projections onto the mean direction (Ley et al. (1975), Mushkudiani (2002), Pandolfo et al. (2018));
- (iii) kernel-based methods: (Hall et al. (1987), García-Portugués et al. (2013), Amiri et al. (2017), Pham Ngoc (2019), Di Marzio et al. (2019), Boente et al. (2014));
- (iv) rank-based methods (Ley et al. (2013, 2017), Verdebout (2017)).

To be valid, however, most of these tests require rotational symmetry (the ranks in (iv) are related to latitudes). Although extremely restrictive, that assumption of rotational symmetry is pervasive in the analysis of directional data. The reason is that it actually reduces the  $(d-1)$ -dimensional problem to a univariate one:<sup>11</sup> the latitudes (dimension 1) indeed are minimal sufficient while the (hyper)longitudes (dimension  $d-2$ ) are ancillary and can be dropped. Distribution functions, quantile functions, and ranks, then, with minor adjustments, are those of the latitudes. While making life much easier, however, dropping the (hyper)longitudes induces, in case the assumption of rotational symmetry is violated, a very significant loss of information—the larger  $d$ , the less plausible rotational symmetry, and the larger the potential loss. For instance, when based on latitudes only, a Goodness-of-Fit test cannot detect alternatives under which  $P \neq P_0$  but  $P^{\text{latitude}} = P_0^{\text{latitude}}$ , no matter how different  $P^{\text{longitude}}$  and  $P_0^{\text{longitude}}$ .

<sup>11</sup> The assumption of rotational symmetry, in that respect, plays in  $\mathcal{S}_{d-1}$  the same role as the (equally restrictive) assumption of elliptical symmetry in  $\mathbb{R}^d$ .

Ideally, Goodness-of-Fit tests for directional data that do not require any assumption such as rotational symmetry should be based on a concept of distribution function with domain the (hyper)sphere enjoying all the properties that make the distribution function  $F$  of a univariate  $P$  over  $\mathbb{R}$  the ideal tool in the univariate setting:

- (a)  $F$  entirely characterizes  $P$ ;
- (b) if  $Z \sim P$ , then  $F(Z)$  is distribution-free ( $F(Z) \sim U_{[0,1]}$ );
- (c) the empirical version  $F^{(n)}$  of  $F$  satisfies the Glivenko-Cantelli property: if  $Z_1^{(n)}, \dots, Z_n^{(n)}$  i.i.d. with distribution function  $F$ ,

$$\max_{1 \leq i \leq n} \left| F^{(n)}(Z_i^{(n)}) - F(Z_i^{(n)}) \right| \longrightarrow 0 \text{ a.s. as } n \rightarrow \infty.$$

In Hallin et al. (2022), we show how such a concept can be based on measure transportation ideas. Measure transportation on Polish spaces and, more particularly, on Riemannian manifolds, is a well-studied subject; see, e.g., results by Rüschendorf (1996), McCann (2001), Ambrosio and Pratelli (2003), Pratelli (2008), Schachermayer and Teichmann (2008), and Chap. 5 of Villani (2009).

Skipping mathematical details, these results imply (see Proposition 2.1 in Hallin et al. (2022)) that there exists a unique (a.s.) optimal transport map  $\mathbf{F}$  from  $\mathcal{S}_{d-1}$  to  $\mathcal{S}_{d-1}$  such that, denoting by  $U_{d-1}$  the uniform over  $\mathcal{S}_{d-1}$ ,  $\mathbf{F}\#\mathbf{P} = U_{d-1}$ . Optimality here consists in minimizing, over the family of all mappings  $\mathbf{S}$  from  $\mathcal{S}_{d-1}$  to  $\mathcal{S}_{d-1}$  such that  $\mathbf{S}\#\mathbf{P} = U_{d-1}$ , the expected *transportation cost*  $\int_{\mathcal{S}_{d-1}} c(\mathbf{Z}, \mathbf{S}(\mathbf{Z})) dP(\mathbf{Z}) = E_P[c(\mathbf{Z}, \mathbf{S}(\mathbf{Z}))]$  where  $c(\mathbf{x}, \mathbf{y})$  stands for the squared Riemannian distance between  $\mathbf{x}$  and  $\mathbf{y}$  in  $\mathcal{S}_{d-1}$ . Call  $\mathbf{F}$  the (*directional*) *distribution function of  $\mathbf{Z} \sim P$* .

It can be shown that  $\mathbf{F}$  is a homeomorphism from  $\mathcal{S}_{d-1}$  to  $\mathcal{S}_{d-1}$ ; its inverse  $\mathbf{Q} := \mathbf{F}^{-1}$  (the *directional quantile function of  $\mathbf{Z} \sim P$* ) thus exists, is continuous, and such that  $\mathbf{Q}\#U_{d-1} = P$ . That directional distribution function satisfies all the properties expected from a distribution function: (a) and (b) follow from the fact that  $\mathbf{F}\#\mathbf{P} = U_{d-1}$  and  $\mathbf{F}^{-1}\#U_{d-1} = P$ . Turning to the sample, let  $\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_n^{(n)}$  be i.i.d. with distribution  $P$  and directional distribution function  $\mathbf{F}$  and denote by  $\mathfrak{G}^{(n)}$  a “regular” grid of  $n$  points over  $\mathcal{S}_{d-1}$ —such that the uniform discrete over the  $n$  gridpoints converges weakly to  $U_{d-1}$  as  $n \rightarrow \infty$ . An empirical counterpart  $\mathbf{F}^{(n)}$  to  $\mathbf{F}$  can be constructed as the discrete optimal transport from the sample to  $\mathfrak{G}^{(n)}$ —namely,  $\operatorname{argmin}_{\mathbf{T} \in \mathcal{T}} \sum_{i=1}^n c(\mathbf{Z}_i^{(n)}, \mathbf{T}(\mathbf{Z}_i^{(n)}))$  where  $\mathbf{T}$  ranges over the family  $\mathcal{T}$  of all bijections between the sample and the grid  $\mathfrak{G}^{(n)}$ . Proposition 4.1 in Hallin et al. (2022) establishes that  $\mathbf{F}^{(n)}$  enjoys the Glivenko-Cantelli property (c):  $\max_{1 \leq i \leq n} \left\| \mathbf{F}^{(n)}(\mathbf{Z}_i^{(n)}) - \mathbf{F}(\mathbf{Z}_i^{(n)}) \right\| \longrightarrow 0$  as  $n \rightarrow \infty$ .

Now that we have an adequate concept of (empirical) distribution function, the construction of Goodness-of-Fit tests is intuitively straightforward. Let again  $\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_n^{(n)}$  denote an i.i.d. sample with distribution  $P$  and empirical distribution function  $\mathbf{F}^{(n)}$ . Consider the problem of testing  $\mathcal{H}_0 : P = P_0$  where  $P_0$  has

distribution function  $\mathbf{F}_0$ . That test naturally can be based on a distance between  $\mathbf{F}^{(n)}$  and  $\mathbf{F}_0$ .

The test we are proposing is a Cramér-von Mises-type test that rejects the null hypothesis for large values of the test statistic

$$T_n := n^{-1} \sum_{i=1}^n \left\| \mathbf{F}^{(n)}(\mathbf{Z}_i^{(n)}) - \mathbf{F}_0(\mathbf{Z}_i^{(n)}) \right\|^2. \quad (11)$$

As usual with Goodness-of-Fit tests, critical values  $c_\alpha$  such that  $\mathbb{P}[T_n > c_\alpha] = \alpha$  under the null hypothesis are easily obtained via simulations since the null hypothesis is simple; the resulting test has exact size  $\alpha$  and is universally consistent. More precisely, provided that the uniform discrete distribution over the  $n$ -points grid  $\mathfrak{G}^{(n)}$  converges weakly, as  $n \rightarrow \infty$ , to the uniform distribution  $\mathbb{U}_{d-1}$  over  $\mathcal{S}^{d-1}$ ,

- (i)  $T_n = o_{\mathbb{P}}(1)$  as  $n \rightarrow \infty$  under  $\mathcal{H}_0$ , while
- (ii)  $T_n$  converges in probability, as  $n \rightarrow \infty$ , to a strictly positive constant if  $\mathbb{P} \neq \mathbb{P}_0$ .

To the best of our knowledge, this is the first universally consistent Goodness-of-Fit test for directional data. Simulations indicate that it performs equally well as existing procedures against rotationally symmetric alternatives, but substantially better under the non-symmetric ones.

As an illustration, we investigate, via simulations,<sup>12</sup> the finite-sample performance of our optimal-transport-based Goodness-of-Fit test based on (11) for the null hypothesis of uniformity on  $\mathcal{S}^2$ . In Table 1, the size and power of our test (OT) are compared with those of the projected Cramér-Von Mises (PCvM), projected Anderson-Darling (PAD), and projected Rothman (PRt) tests of uniformity recently studied in García-Portugués et al. (2023). Four types of spherical distributions are used for generating i.i.d. sample of size  $n = 400$ :

- (i) the uniform distribution;
- (ii) the von Mises-Fisher (vMF) distribution  $\mathcal{M}_2(\boldsymbol{\theta}, \kappa)$  with location  $\boldsymbol{\theta} = (0, 0, 1)'$  and concentration parameter  $\kappa = 1/8$ ;
- (iii) the tangent von Mises-Fisher (tangent vMF) distribution<sup>13</sup> as defined in García-Portugués et al. (2020);
- (iv) the mixture  $I_{[U \leq 0.5]} \mathbf{Z}_1 + I_{[0.5 < U < 0.75]} \mathbf{Z}_2 + I_{[U \geq 0.75]} \mathbf{Z}_3$  of a tangent von Mises-Fisher  $\mathbf{Z}_1$  and two von Mises-Fisher  $\mathbf{Z}_2$  and  $\mathbf{Z}_3$ , where  $U \sim \mathbb{U}_{[0,1]}$ ,  $\mathbf{Z}_1$ ,  $\mathbf{Z}_2$ , and  $\mathbf{Z}_3$  are mutually independent,  $\mathbf{Z}_1$  is tangent vMF as described in (iii) but with skewness intensity  $\kappa_1 = 0.3$ ,  $\mathbf{Z}_2 \sim \mathcal{M}_2(\boldsymbol{\theta}_2, \kappa_2)$ , and  $\mathbf{Z}_3 \sim$

<sup>12</sup> Simulations prepared by Hang Liu.

<sup>13</sup> The tangent vMF distribution with location  $\boldsymbol{\theta}$ , skewness direction  $\boldsymbol{\mu}$ , and skewness intensity  $\kappa$  is the distribution of  $\mathbf{Z} := V\boldsymbol{\theta} + \sqrt{1 - V^2}\boldsymbol{\Gamma}_\boldsymbol{\theta}\mathbf{U}$ , where  $\boldsymbol{\Gamma}_\boldsymbol{\theta}$  denotes a  $d \times (d - 1)$  semi-orthogonal matrix such that  $\boldsymbol{\Gamma}_\boldsymbol{\theta}\boldsymbol{\Gamma}_\boldsymbol{\theta}' = \mathbf{I}_d - \boldsymbol{\theta}\boldsymbol{\theta}'$  and  $\boldsymbol{\Gamma}_\boldsymbol{\theta}'\boldsymbol{\Gamma}_\boldsymbol{\theta} = \mathbf{I}_{d-1}$ .  $V$  is an absolutely continuous scalar random variable and  $\mathbf{U} \sim \mathcal{M}_2(\boldsymbol{\mu}, \kappa)$  are mutually independent; in the simulation, we set  $\boldsymbol{\theta} = (0, 0, 1)'$ ,  $\boldsymbol{\mu} = (0, 1)'$ ,  $\kappa = 0.2$ , and  $V \sim U_{[-1,1]}$ .



**Table 1.** Rejection frequencies of the OT (our measure-transportation-based GoF test), PCvM (projected Cramér-Von Mises test), PAD (projected Anderson-Darling test), and PRt (projected Rothman test) tests of uniformity over  $\mathcal{S}^2$  under the uniform, vMF, tangent vMF, and mixtures of two vMFs and a tangent vMF;  $N = 1000$  replications, sample size  $n = 400$ .

	OT	PCvM	PAD	PRt
Uniform	0.055	0.047	0.048	0.043
vMF	0.153	0.177	0.178	0.178
tangent vMF	0.773	0.680	0.676	0.683
Mixture of two vMFs and a tangent vMF	0.627	0.520	0.519	0.523

$\mathcal{M}_2(\boldsymbol{\theta}_3, \kappa_3)$  with locations  $\boldsymbol{\theta}_2 = (0, -0.3, \sqrt{0.91})'$  and  $\boldsymbol{\theta}_3 = (0.3, \sqrt{0.66}, 0.5)'$ , respectively, and concentrations  $\kappa_2 = \kappa_3 = 0.3$ .

Table 1 shows the rejection frequencies (at nominal level  $\alpha = 0.05$ ), out of  $N = 1000$  replications, of the OT, PCvM, PAD and PRt tests of uniformity. The critical values for our test are obtained through 2000 Monte Carlo replications. Under the null (uniform spherical distribution), all tests yield rejection frequencies close to  $\alpha = 5\%$ . Under the rotationally symmetric alternative of a von Mises-Fisher distribution, the competitors (PCvM, PAD and PRt tests) are slightly more powerful than our test (OT). However, when the underlying distribution fails to be rotationally symmetric, (the tangent vMF distribution or the mixture distribution), our test significantly outperforms all others.

## Conclusion

Measure transportation methods are providing the first sound concepts of distribution and quantile functions, hence also ranks and signs in a variety of contexts where such concepts so far were unavailable: observations in  $\mathbb{R}^d$  with  $d \geq 2$  (see Hallin et al. (2021), observations on the  $(d-1)$ -sphere  $\mathcal{S}_{d-1}$ , etc. This opened the door to statistical methods that so far were limited to univariate or single-output models. Three examples of such methods are described here—fully distribution-free rank tests for the multivariate two-sample problem, nonparametric multiple-output quantile regression, and Goodness-of-Fit tests for directional data. There are many more, though: rank-based tests for multiple-output regression and MANOVA (Hallin et al. (2021); see also Ghosal and Sen (2019)), rank-based tests and R-estimation for VAR and VARMA time-series models (Hallin et al. (2021, 2023), Hallin and Liu (2023)); distribution-free tests of vector independence (Deb and Sen (2021), Shi et al. (2022)), etc. and further applications are likely to be developed in the future.

**Acknowledgment.** The help of Alberto González-Sanz, Šarká Hudecova, Hang Liu, and Gilles Mordant in the preparation of the figures and the simulations is gratefully acknowledged.

## References

- Agostinelli, C., Romanazzi, M.: Nonparametric analysis of directional data based on data depth. *Environ. Ecol. Stat.* **20**, 253–270 (2013)
- Ambrosio, L., Pratelli, A.: Existence and stability results in the  $L_1$  theory of optimal transportation. In: Ambrosio, L., Caffarelli, L.A., Brenier, Y., Buttazzo, G., Villani, C., Salsa, S. (eds.) *Optimal Transportation and Applications*. LNM, vol. 1813, pp. 123–160. Springer, Heidelberg (2003). [https://doi.org/10.1007/978-3-540-44857-0\\_5](https://doi.org/10.1007/978-3-540-44857-0_5)
- Amiri, A., Thiam, B., Verdebout, T.: On the estimation of the density of a directional data stream. *Scand. J. Stat.* **44**, 249–267 (2017)
- del Barrio, E., González-Sanz, A., Hallin, M.: A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *J. Multivariate Anal.* **180**, 104671 (2020)
- del Barrio, E., González-Sanz, A., Hallin, M.: Nonparametric multiple-output center-outward quantile regression (2022). <https://doi.org/10.48550/arXiv.2204.11756>
- Boente, G., Rodriguez, D., González-Manteiga, W.: Goodness-of-fit test for directional data. *Scand. J. Stat.* **41**, 259–275 (2014)
- Chernozhukov, V., Galichon, A., Hallin, M., Henry, M.: Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Stat.* **45**, 223–256 (2017)
- Deb, N., Sen, B.: Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Am. Stat. Assoc.* (2021). <https://doi.org/10.1080/01621459.2021.1923508>
- Di Marzio, M., Fensore, S., Panzera, A., Taylor, C.C.: Kernel density classification for spherical data. *Stat. Probab. Lett.* **144**, 23–29 (2019)
- Dua, D., Graff, C.: UCI machine learning repository (2017)
- Figalli, A.: On the continuity of center-outward distribution and quantile functions. *Nonlinear Anal.* **177**, 413–421 (2018)
- Galichon, A.: *Optimal Transport Methods in Economics*. Princeton University Press, Princeton (2016)
- García-Portugués, E., Crujeiras, R.M., González-Manteiga, W.: Kernel density estimation for directional-linear data. *J. Multivar. Anal.* **121**, 152–175 (2013)
- García-Portugués, E., Navarro-Esteban, P., Cuesta-Albertos, J.A.: On a projection-based class of uniformity tests on the hypersphere. *Bernoulli* **29**, 181–204 (2023)
- García-Portugués, E., Paindaveine, D., Verdebout, T.: On optimal tests for rotational symmetry against new classes of hyperspherical distributions. *J. Am. Stat. Assoc.* **115**, 1873–1887 (2020)
- Ghosal, P., Sen, B.: Multivariate ranks and quantiles using optimal transport: consistency, rates, and nonparametric testing. *Ann. Stat.* (2019, to appear)
- Hall, P., Watson, G.S., Cabrera, J.: Kernel density estimation with spherical data. *Biometrika* **74**, 751–762 (1987)
- Hallin, M.: On distribution and quantile functions, ranks and signs in  $\mathbb{R}^d$ : a measure transportation approach (2017). <https://ideas.repec.org/p/eca/wpaper/2013-258262.html>.
- Hallin, M.: Measure transportation and statistical decision theory. *Ann. Rev. Stat. Appl.* **9**, 401–424 (2022)
- Hallin, M., del Barrio, E., Cuesta-Albertos, J., Matrán, C.: Center-outward distribution and quantile functions, ranks, and signs in  $\mathbb{R}^d$ : a measure transportation approach. *Ann. Stat.* **49**, 1139–1165 (2021)
- Hallin, M., Hlubinka, D., Hudecová, Š: Fully distribution-free center-outward rank tests for multiple-output regression and MANOVA. *J. Am. Stat. Assoc.* (2022a, to appear). <http://arxiv.org/abs/2007.15496>

- Hallin, M., La Vecchia, D., Liu, H.: Center-outward R-estimation for semiparametric VARMA models. *J. Am. Stat. Assoc.* **117**, 925–938 (2021)
- Hallin, M., La Vecchia, D., Liu, H.: Rank-based testing for semiparametric VAR models: a measure transportation approach. *Bernoulli* **29**, 229–273 (2023)
- Hallin, M., Liu, H.: Center-outward rank- and sign-based VARMA portmanteau tests: Chitturi, Hosking, and Li-McLeod revisited. *Econometrics Stat.* (2023, to appear). <http://arxiv.org/abs/2208.12143>
- Hallin, M., Liu, H., Verdebout, T.: Nonparametric measure-transportation-based methods for directional data (2022). <https://ideas.repec.org/p/eca/wpaper/2013-344268.html>
- Hallin, M., Lu, Z., Paindaveine, D., Šiman, M.: Local bilinear multiple-output quantile/depth regression. *Bernoulli* **21**, 1435–1466 (2015)
- Hallin, M., Mordant, G., Segers, J.: Multivariate goodness-of-fit tests based on Wasserstein distance. *Electron. J. Stat.* **15**, 1328–1371 (2021)
- Hallin, M., Mordant, G.: On the finite-sample performance of measure-transportation-based multivariate rank tests. In: Yi, M., Nordhausen, K. (eds.) *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*, pp. 87–119. Springer, Berlin (2023). [arXiv:2111.04705](https://arxiv.org/abs/2111.04705)
- Hallin, M., Šiman, M.: Multiple-output quantile regression. In: Koenker, R., Chernozhukov, V., He, X., Peng, L. (eds.) *Handbook of Quantile Regression*, pp. 185–207. CRC Press, Boca Raton (2018)
- Hallin, M., Paindaveine, D., Šiman, M.: Multivariate quantiles and multiple-output regression quantiles: from  $L_1$  optimization to halfspace depth [with Discussion and Rejoinder]. *Ann. Stat.* **38**, 635–703 (2010)
- Jammalamadaka Rao, S., Meintanis, S., Verdebout, T.: On new Sobolev tests of uniformity on the circle with extension to the sphere. *Bernoulli* **26**, 2226–2252 (2020)
- Jupp, P.E.: Sobolev tests of goodness of fit of distributions on compact Riemannian manifolds. *Ann. Stat.* **33**, 2957–2966 (2005)
- Jupp, P.E.: Data-driven Sobolev tests of uniformity on compact Riemannian manifolds. *Ann. Stat.* **36**, 1246–1260 (2008)
- Koenker, R.: *Quantile Regression*. Econometric Society Monographs, Cambridge University Press, Cambridge (2005)
- Koenker, R., Bassett, G.: Regression quantiles. *Econometrica* **46**, 33–50 (1978)
- Koenker, R., Chernozhukov, V., He, X., Peng, L. (eds.): *Handbook of Quantile Regression*. CRC Press (2018)
- Kong, L., Mizera, I.: Quantile tomography: using quantiles with multivariate data. *Stat. Sin.* **22**, 1589–1610 (2012)
- Lehmann, E.L.: *Nonparametrics: Statistical Methods Based on Ranks*. Mc Graw-Hill, New York (1975)
- Ley, C., Swan, Y., Verdebout, T.: Efficient ANOVA for directional data. *Ann. Inst. Stat. Math.* **69**, 39–62 (2017)
- Liu, R.Y., Singh, K.: Ordering directional data: concepts of data depth on circles and spheres. *Ann. Stat.* **20**, 1468–1484 (1992)
- McCann, R.J.: Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80**, 309–323 (1995)
- McCann, R.: Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.* **11**, 589–608 (2001)
- Mushkudiani, N.A.: Small nonparametric tolerance regions for directional data. *J. Stat. Plann. Inference* **100**, 67–80 (2002)
- Panaretos, V., Zemel, Y.: Statistical aspects of Wasserstein distances. *Ann. Rev. Stat. Appl.* **6**, 405–31 (2019)

- Pandolfo, G., Paindaveine, D., Porzio, G.C.: Distance-based depths for directional data. *Can. J. Stat.* **46**, 593–609 (2018)
- Pham Ngoc, T.M.: Adaptive optimal kernel density estimation for directional data. *J. Multivar. Anal.* **173**, 248–267 (2019)
- Pratelli, A.: On the sufficiency of  $c$ -cyclical monotonicity for optimality of transport plans. *Math. Z.* **258**, 677–690 (2008)
- Rousseeuw, P.J., Struyf, A.: Characterizing angular symmetry and regression symmetry. *J. Stat. Plann. Inference* **122**, 161–173 (2004)
- Rüschendorf, L.: On  $c$ -optimal random variables. *Stat. Probab. Lett.* **27**, 267–270 (1996)
- Schachermayer, W., Teichmann, J.: Characterization of optimal transport plans for the Monge-Kantorovich problem. *Proc. Am. Math. Soc.* **137**, 519–529 (2008)
- Shi, H., Hallin, M., Drton, M., Han, F.: On universally consistent and fully distribution-free rank tests of vector independence. *Ann. Stat.* **50**, 1933–1959 (2022)
- Verdebout, T.: On the efficiency of some rank-based test for the homogeneity of concentrations. *J. Stat. Plann. Inference* **191**, 101–109 (2017)
- Villani, C.: *Optimal Transport: Old and New*. Grundlehren der Mathematischen Wissenschaften, vol. 338. Springer, Berlin and Heidelberg (2009). <https://doi.org/10.1007/978-3-540-71050-9>
- Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945)



# Extending the A Priori Procedure for Estimating Location Parameter Under Multivariate Skew Normal Settings

Ziwei Ma<sup>1</sup>, Tonghui Wang<sup>2(✉)</sup>, S. T. Boris Choy<sup>3</sup>, Zheng Wei<sup>4</sup>,  
and Xiaonan Zhu<sup>5</sup>

<sup>1</sup> Department of Mathematics, The University of Tennessee at Chattanooga,  
Chattanooga, USA

ziwei-ma@utc.edu

<sup>2</sup> Department of Mathematical Sciences, New Mexico State University,  
Las Cruces, USA

twang@nmsu.edu

<sup>3</sup> Discipline of Business Analytics, The University of Sydney, Sydney, Australia  
boris.choy@sydney.edu.au

<sup>4</sup> Department of Mathematics and Statistics,  
Texas A& M University - Corpus Christi, Corpus Christi, USA

zheng.wei@tamucc.edu

<sup>5</sup> Department of Mathematics, University of North Alabama, Florence, USA  
xzhu7@una.edu

**Abstract.** In this work, a multivariate version of the a priori procedure (APP) for estimating the vector of location parameters under skew normal assumptions is studied, in which the necessary sample size to meet the given precision and the level of confidence is provided. The APP is a useful tool for researchers to determine the necessary sample size to reach goals pertaining to precision and confidence level simultaneously. Previous researchers focused mostly on univariate and bivariate cases. The present work addresses the lack of applications under the umbrella of the multivariate skew normal distributions. In addition to derivations of relevant equations, there is a link to a free and user-friendly computer program. Finally, we present computer simulations and a real data example to support our main results.

**Keywords:** A priori procedure · Multivariate skew-normal distribution · confidence · vector of location parameters

## 1 Introduction

Recently, the a priori procedure (APP) initially proposed by Trafimow [7] and further developed by Trafimow and colleagues in both normal and skew normal

populations [8], such as estimating differences between means for both dependence and independent cases [9–11]. However, all of those works focused on the univariate and bivariate cases. In many practical problems, there are multiple values observed for each subject so that the extension the APP procedures to include multivariate cases is needed.

We consider the distribution of observations having multivariate skew normal (MSN) distribution which includes multivariate normal distribution with an extra vector of skewness (or shape) parameters [1]. A  $k$ -dimensional random vector  $\mathbf{Y}$  follows a skew normal distribution with the vector of location parameters  $\boldsymbol{\mu} \in \mathfrak{R}^k$ , matrix of scale parameters  $\Sigma$  (a  $k \times k$  positive definite matrix), and the vector of skewness parameters  $\boldsymbol{\lambda} \in \mathfrak{R}^k$ , denoted by  $\mathbf{Y} \sim SN_k(\boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda})$ , if its probability density function (pdf) is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = 2\phi_k(\mathbf{y}; \boldsymbol{\mu}, \Sigma) \Phi\left(\boldsymbol{\lambda}'\Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\right), \quad \mathbf{y} \in \mathfrak{R}^k, \quad (1)$$

where  $\phi_k(\cdot)$  represents the pdf of  $k$ -dimensional normal distribution and  $\Phi(\cdot)$  represents the cumulative distribution function (cdf) of univariate standard normal distribution. Note that the distribution of  $\mathbf{Y}$  is reduced to  $SN_k(\boldsymbol{\lambda})$  if  $\boldsymbol{\mu} = \mathbf{0}$  and  $\Sigma = I_k$ , the  $k$ -dimensional identity matrix. Also when  $\boldsymbol{\lambda} = \mathbf{0}$ ,  $\mathbf{Y}$  is multivariate normal distributed,  $N_k(\boldsymbol{\mu}, \Sigma)$ . We focus on determining the minimum sample size required for estimating the location parameter  $\boldsymbol{\mu}$  with pre-specified precision and level of confidence using the APP.

This paper is organized as follows. In Sect. 2, the matrix variate skew normal distribution is introduced which describes the joint distribution of the random sample matrix. The sampling distributions are presented as well. In Sect. 3, the main result, minimum sample size required for estimating the vector of location parameters  $\boldsymbol{\mu}$ , is derived. For an illustration of the main results, a simulation study is conducted for various values of dimensions in Sect. 4. A real data example is given in Sect. 5 and the conclusion remarks are listed in Sect. 6.

## 2 Review of Matrix Variate SN Distributions and Sampling Distributions

Let  $M_{n \times k}$  be the set of all  $n \times k$  matrices over the real field  $\mathfrak{R}$  and  $\mathfrak{R}^n = M_{n \times 1}$ . The identity matrix is denoted as  $I_n \in M_{n \times n}$ , the constant vector  $(1, \dots, 1)' \in \mathfrak{R}^n$  is denoted as  $\mathbf{1}_n$ , and  $\bar{J}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ . For  $B = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n)' \in M_{n \times k}$  with  $\mathbf{b}_i \in \mathfrak{R}^k$ , let  $\text{Vec}(B) = (\mathbf{b}'_1, \mathbf{b}'_2, \dots, \mathbf{b}'_n)' \in \mathfrak{R}^{nk}$ . For positive definite matrix  $T \in M_{n \times n}$ , we use  $T^{-1}$  to denote the inverse. For  $B \in M_{m \times n}$ ,  $C \in M_{n \times p}$ , we use  $B \otimes C$  to denote the Kronecker product of  $B$  and  $C$ . Through this paper,  $N(0, 1)$  represents the standard normal distribution and bold phase letters represent vectors.

**Definition 1.** The  $n \times k$  random matrix  $Y$  is said to have a skew-normal matrix distribution with location matrix  $M$ , scale matrix  $V \otimes \Sigma$  and skewness parameter matrix  $\boldsymbol{\gamma} \otimes \boldsymbol{\lambda}'$ , denoted by  $Y \sim SN_{n \times p}(M, V \otimes \Sigma, \boldsymbol{\gamma} \otimes \boldsymbol{\lambda}')$ , if  $\mathbf{y} \equiv \text{Vec}(Y) \sim$

$SN_{np}(\boldsymbol{\mu}_0, V \otimes \Sigma, \boldsymbol{\gamma} \otimes \boldsymbol{\lambda})$ , where  $M \in M_{n \times p}$ ,  $V \in M_{n \times n}$ ,  $\boldsymbol{\mu}_0 = \text{Vec}(M)$ ,  $\boldsymbol{\gamma} \in \mathfrak{R}^n$  and  $\boldsymbol{\lambda} \in \mathfrak{R}^p$ .

Suppose that  $Y \in M_{n_1 \times p}$  is the sample matrix such that

$$Y \sim SN_{n \times p}(\mathbf{1}_n \otimes \boldsymbol{\mu}'_0, I_n \otimes \Sigma, \mathbf{1}_n \otimes \boldsymbol{\lambda}'). \tag{2}$$

**Definition 2.** (Wang et al. [12] and Ye et al. [13]). Let  $\mathbf{X} \sim SN_m(\boldsymbol{\nu}, I_m, \boldsymbol{\lambda})$ . The distribution of  $\mathbf{X}'\mathbf{X}$  is defined as the **noncentral skew chi-square distribution** with degrees of freedom  $m$ , the noncentral parameter  $\xi = \boldsymbol{\nu}'\boldsymbol{\nu}$ , and the skewness parameters  $\delta_1 = \boldsymbol{\lambda}'\boldsymbol{\nu}$  and  $\delta_2 = \boldsymbol{\lambda}'\boldsymbol{\lambda}$ , denoted by  $\mathbf{Y}'\mathbf{Y} \sim S\chi_m^2(\xi, \delta_1, \delta_2)$ . In particular, if  $\delta_1 = 0$ , then  $\mathbf{X}'\mathbf{X} \sim \chi_m^2(\xi)$ , which is free of the vector of skewness parameter.

The following lemmas will be used the proof our main results.

**Lemma 1.** (Ma et al. [4]). *Let  $Y \sim SN_{n \times p}(\mathbf{1}_n \otimes \boldsymbol{\mu}'_0, I_n \otimes \Sigma, \mathbf{1}_n \otimes \boldsymbol{\lambda}')$ . Then the following results hold.*

- (a)  $\bar{Y} = (\mathbf{1}'_n Y/n)' \sim SN_p(\boldsymbol{\mu}_0, \Sigma/n, \sqrt{n}\boldsymbol{\lambda})$ .
- (b)  $(n-1)S = Y'(I_n - \bar{J}_n)Y \sim W_p(n-1, \Sigma)$ .
- (c)  $\bar{Y}$  and  $S$  are independent.
- (d)  $\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{Y} - \boldsymbol{\mu}_0) \sim SN_p(\mathbf{0}, I_p\sqrt{n}\boldsymbol{\lambda})$  so that  $n(\bar{Y} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\bar{Y} - \boldsymbol{\mu}_0) \sim \chi_p^2$ .

Here  $W_p(n-1, \Sigma)$  represents the  $p$ -dimensional Wishart distribution with mean  $\Sigma$  and  $n-1$  degrees of freedom.

**Lemma 2.** (Mardia et al. [5], Theorem 3.4.7). *If  $M \sim W_p(m, \Sigma)$ ,  $m > p$ , then the ratio  $\mathbf{a}'\Sigma^{-1}\mathbf{a}/\mathbf{a}'M^{-1}\mathbf{a} \sim \chi_{m-p+1}^2$  for any fixed  $p$ -vector  $\mathbf{a}$ .*

### 3 Required Sample Size Needed for Estimating the Location Parameter

In this section, we derive the main results of determining the desired sample size for location parameter  $\boldsymbol{\mu}_0$ . We employ the Hotelling's  $T^2$

$$T^2 = n(\bar{Y} - \boldsymbol{\mu}_0)'S^{-1}(\bar{Y} - \boldsymbol{\mu}_0). \tag{3}$$

The Hotelling  $T^2$  is commonly used to test the hypothesis of the location parameter vector, say  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ . The distribution of  $T^2$  has been derived in [3] which has noncentral skew F distribution. The definition and pdf are given below.

**Definition 3.** Assume that  $U_1 \sim S\chi_{n_1}^2(\xi, \delta_1, \delta_2)$ ,  $U_2 \sim \chi_{n_2}^2$ , and  $U_1$  and  $U_2$  are independent. The distribution of  $F = \frac{U_1/n_1}{U_2/n_2}$  is called the **noncentral skew F distribution** with degrees of freedom  $n_1$  and  $n_2$ , the noncentral parameter  $\xi$ , and the skewness parameters  $\delta_1$  and  $\delta_2$ , denoted by  $F \sim SF_{n_1, n_2}(\xi, \delta_1, \delta_2)$ . The pdf of  $F$  is

$$f_F(x; \xi, \delta_1, \delta_2) = \int_0^\infty \int_0^\infty \frac{n_2(u+v)}{n_1x^2} g(u) h(v) k\left(\frac{n_2(u+v)}{n_1x}\right) dudv, \quad (4)$$

where

$$g(s) = s^{-1/2} \left\{ \phi\left(\sqrt{s} - \frac{\delta_1}{\sqrt{\delta_2}}\right) \Phi\left(\sqrt{\delta_2}\left(\sqrt{s} - \frac{\delta_1}{\sqrt{\delta_2}}\right)\right) + \phi\left(\sqrt{s} + \frac{\delta_1}{\sqrt{\delta_2}}\right) \Phi\left(\sqrt{\delta_2}\left(-\sqrt{s} - \frac{\delta_1}{\sqrt{\delta_2}}\right)\right) \right\},$$

is the pdf of  $S\chi_1^2\left(\frac{\delta_1^2}{\delta_2}, \delta_1, \delta_2\right)$ , and  $h(\cdot)$  and  $k(\cdot)$  are pdfs of  $\chi_{n_1-1}^2\left(\xi - \frac{\delta_1^2}{\delta_2}\right)$  and  $\chi_{n_2}^2$ , respectively.

**Lemma 3.** *Under the assumptions in Lemma 1,  $n(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0)$  has non-central skew  $\chi_p^2(\xi, \delta_1, \delta_2)$  where  $\mu_* = \sqrt{n}\Sigma^{-1/2}(\mu - \mu_0)$ ,  $\xi = n\mu_*'\mu_*$ ,  $\delta_1 = n\mu_*'\lambda$  and  $\delta_2 = n\lambda'\lambda$ .*

**Lemma 4.** *Let the sample matrix  $Y \sim SN_{n \times p}(\mathbf{1}_n \otimes \mu_0', I_n \otimes \Sigma, \mathbf{1}_n \otimes \lambda')$ , and  $\bar{Y}$ ,  $S$  and  $T^2$  are defined in Lemma 1 (a), (b) and (3), respectively. Then*

$$T^2 \sim \frac{(n-1)p}{n-p} SF_{p, n-p}(\xi, \delta_1, \delta_2). \quad (5)$$

where  $\mu_* = \sqrt{n}\Sigma^{-1/2}(\mu - \mu_0)$ ,  $\xi = n\mu_*'\mu_*$ ,  $\delta_1 = n\mu_*'\lambda$  and  $\delta_2 = n\lambda'\lambda$ .

Based on the distribution of  $T^2$ , we obtain the following results by applying APP.

**Theorem 1.** *Suppose that a sample matrix  $Y$  follows the distribution defined by Lemma 1, and  $\Sigma$  is unknown and  $\lambda$  is known. Let  $c$  be the confidence level and  $f$  be the precision. The required sample size  $n$  can be obtained by solving*

$$P(\|\Sigma^{-1/2}(\bar{Y} - \mu_0)\| \leq \sqrt{2pf}) = c. \quad (6)$$

where

$$f^2 = \frac{n-1}{2n} SF_{p, n-p}^{-1}(c; \xi, \delta_1, \delta_2). \quad (7)$$

**Proof.** To derive Eq. (6), we square  $\|\Sigma^{-1/2}(\bar{Y} - \mu_0)\|$ , the deviation between parameter vector  $\mu_0$  and its estimator  $\bar{Y}$ , which can be written as a quadratic form, i.e.  $\|\sqrt{n}\Sigma^{-1/2}(\bar{Y} - \mu_0)\|^2 = n(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0)$ . In one hand, the distribution of  $n(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0)$  is  $S\chi_p^2$ . On the other hand, we let  $\omega = \frac{(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0)}{(\bar{Y} - \mu_0)'\Sigma^{-1}(\bar{Y} - \mu_0)}$  which is the variation of the estimator. The distribution of  $\omega$  is proportional to  $\chi_{n-p}^2$  derived by Lemma 2. By APP, we consider

$$P\left(\|\Sigma^{-1/2}(\bar{Y} - \mu_0)\| \leq \sqrt{2pf}\right) = c,$$



which is equivalent to

$$P\left(\frac{\|\Sigma^{-1/2}(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)\|}{\sqrt{\omega}} \leq \frac{\sqrt{2pf}}{\sqrt{\omega}}\right) = c.$$

Then, we square both sides of the inequality, which is

$$P\left(T^2 \leq \frac{2npf^2}{\omega}\right) = c. \tag{8}$$

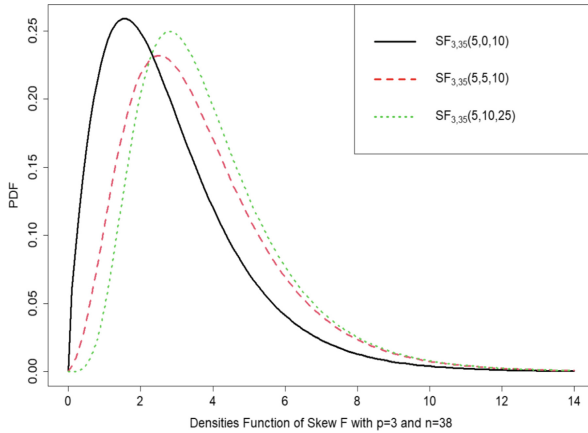
Since  $\omega \sim \chi_{n-p}^2$ , the Eq. (8) can be reduced to

$$P\left(T^2 \leq \frac{2npf^2}{n-p}\right) \approx c$$

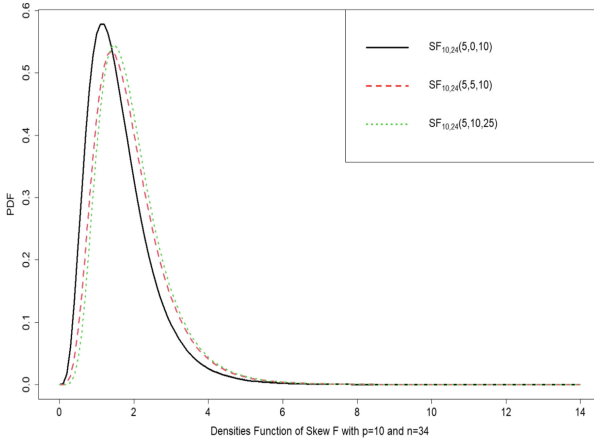
by strong law of large numbers  $\omega$  approaching its expected value  $n-p$  under  $H_0$  in probability. Therefore, the desired result Eq. (6) can be derived by applying Lemma 3.1 and straightforward computation.  $\square$

**Remark.** The required sample size  $n$  will be determined by the smallest integer which makes the right side closest to the left side.

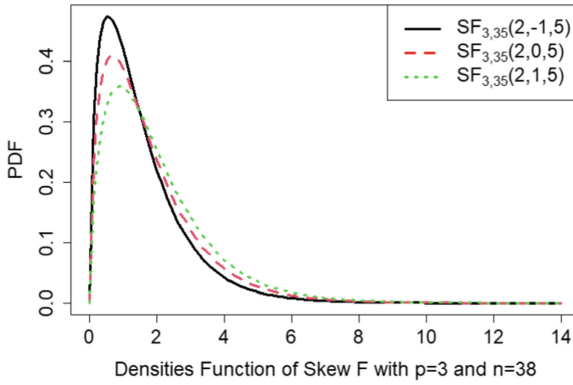
We present Fig. 1–4 to illustrate the pdf of the test statistics  $T^2$  in general which has non central skew F distribution defined in [14].



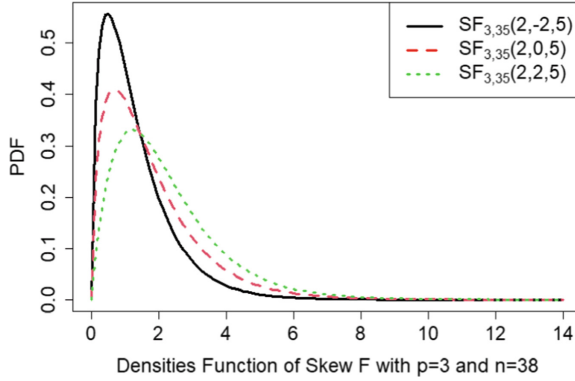
**Fig. 1.** Density curves of  $SF_{3,35}(5,0,10)$  (solid black curve),  $SF_{3,35}(5,5,10)$  (dashed red curve) and  $SF_{3,35}(5,10,25)$  (dashed green curve).



**Fig. 2.** Density curves of  $SF_{10,24}(5, 0, 25)$  (solid black curve),  $SF_{10,24}(5, 10, 25)$  (dashed red curve) and  $SF_{10,24}(5, -10, 25)$  (dot green curve).



**Fig. 3.** Density curves of  $SF_{3,35}(2, -1, 5)$  (solid black curve),  $SF_{3,35}(2, 0, 5)$  (dashed red curve) and  $SF_{3,35}(2, 1, 5)$  (dashed green curve).



**Fig. 4.** Density curves of  $SF_{3,35}(2, -2, 5)$  (solid black curve),  $SF_{3,35}(2, 0, 5)$  (dashed red curve) and  $SF_{3,35}(2, 2, 5)$  (dashed green curve).

**Corollary 1.** When  $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ , the distribution of  $T^2$  is proportional to  $F_{p,n-p}$ . Thus, for given confidence level  $c$  and precision  $f$  under the same assumption in Theorem 1. The required sample size  $n$  can be obtained by solving the equation

$$P\left(\|\Sigma^{-1/2}(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)\| \leq \sqrt{2pf}\right) = c. \tag{9}$$

where

$$f^2 = \frac{n-1}{2n} F_{p,n-p}^{-1}(c). \tag{10}$$

**Remark.** This result is useful in the hypothesis test procedure when the null hypothesis  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ . The required sample size will be calculated based on this result when the confidence level and precision are given. A Shiny app [6] page is available for readers who are interested in applying derived results. The link is <https://ziweimath.shinyapps.io/APPMSN/> which also has a function to plot the pdf of noncentral skew F distribution.

## 4 Simulation Study

In this section, we conduct simulations to verify the derived theoretical results in Sect. 3. For given confidence level  $c = 0.9$  and  $0.95$ , Table 1 shows the relative coverage frequencies of various combinations of precision  $f$ , dimension  $p$ .

**Table 1.** The required sample size and relative frequencies (r.f.) related to the confidence  $c = 0.95, 0.90$  and precision  $f = 0.15, 0.20$  and  $0.25$  for dimension  $p = 3, 5$  and  $10$  when  $(\xi, \delta_1, \delta_2) = (2, -1, 2)$ .

$(\xi, \delta_1, \delta_2) = (2, -1, 2)$		$p = 3$		$p = 5$		$p = 10$	
	$f$	$n$	r.f	$n$	r.f	$n$	r.f
$c = 0.95$	0.15	138	0.9495	129	0.9501	115	0.9582
	0.20	67	0.9469	61	0.9535	58	0.9426
	0.25	38	0.9503	35	0.9416	32	0.9485
$c = 0.90$	0.15	119	0.8986	103	0.9035	95	0.9049
	0.20	52	0.9005	50	0.9033	48	0.8978
	0.25	28	0.8961	25	0.9022	21	0.9005

**Table 2.** The required sample size and relative frequencies (r.f.) related to the confidence  $c = 0.95, 0.90$  and precision  $f = 0.15, 0.20$  and  $0.25$  for dimension  $p = 3, 5$  and  $10$  when  $(\xi, \delta_1, \delta_2) = (2, 0, 2)$ .

$(\xi, \delta_1, \delta_2) = (2, 0, 2)$		$p = 3$		$p = 5$		$p = 10$	
	$f$	$n$	r.f	$n$	r.f	$n$	r.f
$c = 0.95$	0.15	136	0.9501	119	0.9503	105	0.9520
	0.20	63	0.9482	57	0.9475	54	0.9467
	0.25	38	0.9522	35	0.9520	36	0.9500
$c = 0.90$	0.15	109	0.8970	99	0.8961	93	0.9008
	0.20	51	0.9060	48	0.9038	46	0.8991
	0.25	27	0.9062	23	0.8990	21	0.8973

**Table 3.** The required sample size and relative frequencies (r.f.) related to the confidence  $c = 0.95, 0.90$  and precision  $f = 0.15, 0.20$  and  $0.25$  for dimension  $p = 3, 5$  and  $10$  when  $(\xi, \delta_1, \delta_2) = (2, 1, 2)$ .

$(\xi, \delta_1, \delta_2) = (2, 1, 2)$		$p = 3$		$p = 5$		$p = 10$	
	$f$	$n$	r.f	$n$	r.f	$n$	r.f
$c = 0.95$	0.15	126	0.9481	114	0.9511	101	0.9462
	0.20	60	0.9512	53	0.9497	52	0.9526
	0.25	35	0.9527	34	0.9514	35	0.9537
$c = 0.90$	0.15	104	0.9010	96	0.9024	91	0.9008
	0.20	53	0.8955	47	0.9049	44	0.8962
	0.25	31	0.9014	28	0.8986	25	0.9013

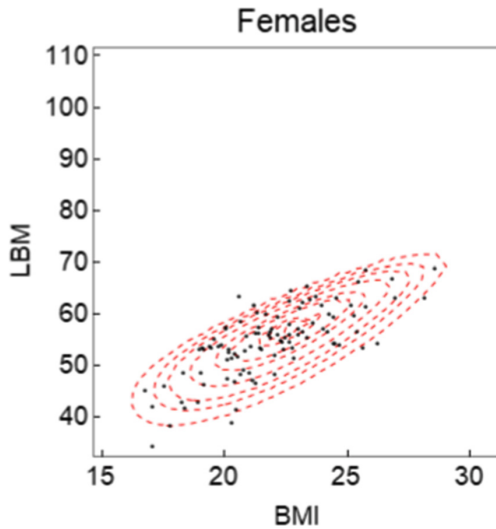
From Table 1, 2 and 3, it shows clearly that the Hotelling  $T^2$  works well on inference location parameters under the MSN setting based on the relative frequencies reported. There is an interesting phenomenon regarding the required sample size. At the same confidence level and precision, the larger dimension  $p$  needs a smaller sample size required to meet the requirement when precision is very high. For example, for  $f = 0.15$ , the required sample sizes are 138 and 129 for  $p = 3$  and  $p = 10$  at confidence level  $c = 0.90$  when  $\xi = 2, \delta_1 = -1, \delta_2 = 2$ . respectively.

### 5 Real Data Example

In this section, the Australian institute of sport (AIS) data [2] is used to demonstrate how to apply APP for determining the required sample size in inference on location parameters for multivariate skew data. We use the variables, body mass index (BMI) and lean body mass (LBM), to illustrate the results. The scatter plots and estimation had been done by Ma et al. [3] which are presented in Table 4 and Fig. 5 as follows.

**Table 4.** Point estimates of skew normal parameters for the female AIS data.

	Females
$\hat{\mu}'$	(23.18, 61.42)
$\hat{\Sigma}$	$\begin{pmatrix} 8.32 & 21.30 \\ 21.30 & 89.96 \end{pmatrix}$
$\hat{\lambda}'$	(0.88, -2.63)



**Fig. 5.** The scatter plots and contour plots for the AIS data of female.

To determine the required sample size from this dataset meeting the specified confidence level and precision, say  $c = 0.9$  and  $f = 0.2$ , applying Eq. 9, we have the smallest sample size  $n = 31$  which meets both confidence level and precision.

## 6 Conclusion

The present work extends an APP for multivariate skew normal setting by employing the generalized Hotelling's  $T^2$ . It provides a useful tool for researchers who need to design an experiment with a given confidence level and precision requirement. Also, a Shiny app is provided for readers who are interested in applying the derived results in this work.

## References

1. Azzalini, A., Valle, A.D.: The multivariate skew-normal distribution. *Biometrika* **83**(4), 715–726 (1996)
2. Cook, R.D., Weisberg, S.: *An Introduction to Regression Graphics*, vol. 405. John Wiley & Sons, Hoboken (2009)
3. Ma, Z., Chen, Y.-J., Wang, T., Liu, J.: Inferences on location parameter in multivariate skew-normal family with unknown scale parameter. *Commun. Stat.-Simul. Comput.* **51**(9), 5465–5481 (2022)
4. Ma, Z., Chen, Y.-J., Wang, T., Peng, W.: The inference on the location parameters under multivariate skew normal settings. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *ECONVN 2019. SCI*, vol. 809, pp. 146–162. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_11](https://doi.org/10.1007/978-3-030-04200-4_11)
5. Mardia, K.V., Kent, J.T., Bibby, J.M.: *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press Inc., New York (1980)
6. RStudio, Inc. *Easy web applications in R* (2013). <http://www.rstudio.com/shiny/>
7. Trafimow, D.: Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. *Educ. Psychol. Meas.* **77**(5), 831–854 (2017)
8. Trafimow, D., MacDonald, J.A.: Performing inferential statistics prior to data collection. *Educ. Psychol. Meas.* **77**(2), 204–219 (2017)
9. Trafimow, D., Wang, C., Wang, T.: Making the a priori procedure work for differences between means. *Educ. Psychol. Meas.* **80**(1), 186–198 (2020)
10. Wang, C., Wang, T., Trafimow, D., Chen, J.: Extending a priori procedure to two independent samples under skew normal settings. *Asian J. Econ. Bank.* **3**(02), 29–40 (2019)
11. Wang, C., Wang, T., Trafimow, D., Myüz, H.A.: Necessary sample sizes for specified closeness and confidence of matched data under the skew normal setting. *Commun. Stat.-Simul. Comput.* **51**(5), 2083–2094 (2022)
12. Wang, T., Li, B., Gupta, A.K.: Distribution of quadratic forms under skew normal settings. *J. Multivar. Anal.* **100**(3), 533–545 (2009)
13. Ye, R., Wang, T., Gupta, A.K.: Distribution of matrix quadratic forms under skew-normal settings. *J. Multivar. Anal.* **131**(00010), 229–239 (2014)
14. Ye, R., Wang, T., Sukparungsee, S., Gupta, A.K.: Tests in variance components models under skew-normal settings. *Metrika* **78**(7), 885–904 (2015). <https://doi.org/10.1007/s00184-015-0532-1>



# A Note on Cournot-Nash Equilibria and Optimal Transport Between Unequal Dimensions

Luca Nenna<sup>1(✉)</sup> and Brendan Pass<sup>2</sup>

<sup>1</sup> Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France  
luca.nenna@universite-paris-saclay.fr

<sup>2</sup> Department of Mathematical and Statistical Sciences, University of Alberta, 632 CAB,  
Edmonton, AB T6G 2G1, Canada  
pass@ualberta.ca

**Abstract.** This note is devoted to study a class of games with a continuum of players for which Cournot-Nash equilibria can be obtained by minimisation of some cost related to Optimal Transport. In particular we focus on the case of an Optimal Transport term between unequal dimension. We also present several numerical simulations.

**Keywords:** Cournot-Nash equilibria · optimal transport · unequal dimensions · nestedness

## 1 Introduction

Since Aumann's seminal works [1, 2], equilibria in games with a continuum of players have received a lot of attention from the Economics and Game Theory communities. Following Mas-Colell's approach [13], we consider a type space  $X \subset \mathbb{R}^m$  endowed with a probability measure  $\mu$ . Each player of type  $x$  has to choose a strategy  $y \in Y$  (where  $Y \subset \mathbb{R}^n$ ) in order to minimise a cost  $\Phi : X \times Y \times \mathcal{P}(Y) \rightarrow \mathbb{R}$  which depends both on his type  $x$  and strategy  $y$  as well as the distribution of strategies  $\nu \in \mathcal{P}(Y)$  resulting from the other players' behaviour. Since the cost depends on other players' strategies only through the distribution  $\nu$ , it is not important who plays a specific strategy, but how many players chose it (i.e. the game is anonymous). Thus, a Cournot-Nash equilibrium is defined as a probability measure  $\gamma \in \mathcal{P}(X \times Y)$ , whose first marginal is  $\mu$ , such that

$$\gamma(\{(x, y) \in X \times Y : \Phi(x, y, \nu) = \min_{z \in Y} \Phi(x, z, \nu)\}) = 1$$

where the strategy distribution  $\nu$  is the second marginal of  $\gamma$ . Notice that if we consider an homogenous population of players (which mean that we do not need a distribution  $\mu$  anymore) then we retrieve exactly the definition of Nash equilibrium.

Let us now focus on the additively separate case where the total cost  $\Phi$  can be written as  $\Phi(x, y, \nu) = c(x, y) + F(\nu, y)$ . It has been recently showed in [6] (see also [4]) that Cournot-Nash equilibria can be obtained by the minimization of a certain functional on

the set of measures on the space of strategies, which actually means that the games considered in [6] belongs to a class of models which have the structure of a potential game (i.e. variational problems). This functional typically involves two terms: an optimal transport cost and a more standard integral functional which may capture both congestion and attractive effects. This variational approach is somehow more constructive and informative (but less general since it requires a separable cost) than the one relying on fixed-point arguments as in [13] but the optimal transport cost term is delicate to handle. It is indeed costly in general to solve numerically an optimal transport problem and compute an element of the subdifferential of the optimal cost as a function of its marginals. However it has been recently introduced (see for instance [3, 11, 12]) a very efficient numerical method relying on an entropic regularization of optimal transport which turned to be useful also to approximate Cournot-Nash equilibria (see [7]), in the case of potential games.

In [6] the authors treat the case where the measures  $\mu$  and  $\nu$  are probabilities on  $X \subset \mathbb{R}^m$  and  $Y \subset \mathbb{R}^n$ , respectively, where  $m = n$ . In this article we want to address the case in which  $m > n$  so that the optimal transport term becomes an *unequal dimensional Optimal Transport term* [8, 9]. In a recent paper [15] we have showed that for a variety of different forms of  $\mathcal{F}[\nu](y)$  a nestedness condition, which is known to yield much improved tractability of the optimal transport problem, holds for any minimizer. Depending on the exact form of the functional, we exploit this to find local differential equations characterizing solutions, prove convergence of an iterative scheme to characterise the solution, and prove regularity results. The aim of this paper is to illustrate some numerical methods to compute the solution in the case of Cournot-Nash equilibria by using the characterization results established in [15].

## Motivation

We want to briefly remark through an example that unequal dimensional Optimal Transport is actually quite natural in the Economics/Game theory framework as the one we are dealing with.

Let think that we have a population of physicians who are characterized by a type  $x$  belonging to  $X \subset \mathbb{R}^m$  where the dimension  $m$  is the number of characteristics: age, gender, university where they get their diploma, their hometown, etc. The probability measure  $\mu$  on  $X$  gives then the distribution of types in the physician population. The main idea of Optimal Transport is to match the physician with a city, for example, where they would like to open their private practice. The cities are characterized by a type  $y$  which belongs to  $Y \subset \mathbb{R}^n$  where  $n$  is now the number of characteristics and  $Y$  is endowed with a probability measure  $\nu$  giving the distribution of types of cities. Notice that the matching between physicians and cities must minimize a given transportation cost  $c(x, y)$  which can indeed model the cost for physician  $x$  to commute to the city  $y$  where he/she has his/her practice. One can also think, in more economical terms, that the cost is of the form  $c(x, y) = -s(x, y)$  where  $s$  is a surplus and in this case the optimal matching will maximize the given surplus. It is actually quite important to highlight now that the number of characteristics  $n$  of the cities can be much smaller than the one for the physician; indeed it is natural to take  $n = 1$  saying that just a characteristic (i.e. the population) of the city is taken into account by the physicians.



*Remark 1 (Notation).* With a slight abuse of notation we will identify the measure with its density throughout all the paper.

## 2 Optimal Transport Between Unequal Dimensions

Given two probability measures  $\mu$  on  $X \subset \mathbb{R}^m$  and  $\nu$  on  $Y \subset \mathbb{R}^n$ , the Monge-Kantorovich problem consists in transporting  $\mu$  onto  $\nu$  so as to optimise a given cost function  $c(x, y)$ . Indeed we look for a probability measure  $\gamma$  on the product space  $X \times Y$  whose marginals are  $\mu$  and  $\nu$  such that it solves the following maximisation problem

$$\mathcal{T}_c(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\gamma(x, y), \tag{1}$$

where  $\Pi(\mu, \nu) := \{\eta \in \mathcal{P}(X \times Y) : \pi_x(\eta) = \mu, \pi_y(\eta) = \nu\}$  and  $\pi_x : X \times Y \rightarrow X$ .

We refer the reader to [18, 19] for results about the characterisation of optimal maps for the case in which  $m = n$ . Before discussing the case  $m > n$ , it is important to highlight that the Monge-Kantorovich problem admits a dual formulation which is useful in order to understand the solution to (1)

$$\mathcal{T}_c^{dual}(\mu, \nu) := \sup_{(u, v) \in \mathcal{U}} \int_X u(x) d\mu(x) + \int_Y v(y) d\nu(y), \tag{2}$$

where  $\mathcal{U} := \{(u, v) \in L^1(\mu) \oplus L^1(\nu) : u(x) + v(y) \geq c(x, y) \text{ on } X \times Y\}$ . The optimal  $u, v$  are called Kantorovich potentials. The interesting fact is that  $\mathcal{T}_c = \mathcal{T}_c^{dual}$ .

Under mild conditions (for instance, if  $c$  is differentiable, as we are assuming here,  $X$  connected and  $\bar{\mu}(x) > 0$  for all  $x \in X$  [18]), there is a unique solution  $(u, v)$  to the dual problem, up to the addition  $(u, v) \mapsto (u + C, v - C)$  of a constant, known as the Kantorovich potentials and these potentials are  $c$ -concave; that is, they satisfy

$$u(x) = v^c(x) := \min_{y \in Y} [c(x, y) - v(y)], \quad v(y) = u^c(y) := \min_{x \in X} [c(x, y) - u(x)],$$

In particular if both  $\nu$  and  $\mu$  are absolutely continuous, then the potential  $v$  satisfies the Monge-Ampere type equation almost everywhere [14]<sup>1</sup>:

$$v(y) = \int_{\partial^c v(y)} \frac{\det(D_{yy}^2 c(x, y) - D^2 v(y))}{\sqrt{|\det(D_{yx}^2 c D_{xy}^2 c)(x, y)|}} \mu(x) d\mathcal{H}^{m-n}(x), \tag{3}$$

where  $\partial^c v(y) := \{x : u(x) + v(y) = c(x, y)\}$ . In general, this is a *non-local* differential equation for  $v(y)$ , since the domain of integration  $\partial^c v(y)$  is defined using the values of  $\nu$  and  $u = v^c$  throughout  $Y$ ; however, when the model satisfies the generalized nestedness condition (namely  $\partial^c v(y) = X_-(y, Dv(y))$ , where  $X_-(y, p) = \{x \in X : D_y c(x, y) = p\}$ , see [15][Definition 1] for more details), it reduces to the local Eq. [14]:

$$v(y) = \int_{X_-(y, Dv(y))} \frac{\det(D_{yy}^2 c(x, y) - D^2 v(y))}{\sqrt{|\det(D_{yx}^2 c D_{xy}^2 c)(x, y)|}} \mu(x) d\mathcal{H}^{m-n}(x). \tag{4}$$

---

<sup>1</sup> Note that, here and below, our notation differs somewhat from [9] and [14], since we have adopted the convention of minimizing, rather than maximizing, in (1).

### 2.1 Multi-to One-Dimensional Optimal Transport

We consider now the optimal transport problem in the case in which  $m > n = 1$  (for more details we refer the reader to [9]). Let us define the super-level set  $X_{\geq}(y, k)$  as follows

$$X_{\geq}(y, k) := \{x \in X : \partial_y c(x, y) \geq k\},$$

and its strict variant  $X_{>}(y, k) := X_{\geq}(y, k) \setminus X_{=}(y, k)$ . In order to build an optimal transport map  $T$  we take the unique level set splitting the mass proportionately with  $y$ ; that is  $k(y)$  such that

$$\mu(X_{\geq}(y, k(y))) = \nu((-\infty, y]), \tag{5}$$

then we set  $y = T(x)$  for all  $x$  which belongs to  $X_{=}(y, k(y))$ . Notice that if  $x \in X_{=}(y_0, k(y_0)) \cap X_{=}(y_1, k(y_1))$  then the map  $T$  is not well-defined. In order to avoid such a degenerate case we ask that the model  $(c, \mu, \nu)$  satisfies the following property

**Definition 1 (Nestedness  $n = 1$ ).** The model  $(c, \mu, \nu)$  is nested if

$$\forall y_0, y_1 \text{ with } y_1 > y_0, \nu([y_0, y_1]) > 0 \implies X_{\geq}(y_0, k(y_0)) \subset X_{>}(y_1, k(y_1)).$$

Thus, if the model  $(c, \mu, \nu)$  is nested then [9][Theorem 4] assures that  $\gamma_T = (\text{id}, T)_{\#}\mu$ , where the map  $T$  is built as above, is the unique maximiser of (1) on  $\Pi(\mu, \nu)$ . Moreover, the optimal potential  $v(y)$  is given by  $v(y) = \int^y k(t) dt$ .

We remark that nestedness depends on all the data  $(c, \mu, \nu)$  of the problem, in the following we give a condition which can ensure nestedness when only  $c$  and  $\mu$  are fixed.

Consider now the equality  $\nu([-\infty, y]) = \mu(X_{\leq}(y, k(y)))$ , then differentiating it gives the following equation

$$v(y) = \int_{X_{=}(y, k(y))} \frac{D_{yy}^2 c(x, y) - k'(y)}{|D_{xy}^2 c(x, y)|} \mu(x) d\mathcal{H}^{m-1}(x), \tag{6}$$

which can be viewed as the multi-to-one counterpart of the Monge-Ampère equation in classical Optimal Transport theory (note also that it is exactly the  $n = 1$  case of (4)).

We now briefly recall some results of [15] which guarantee the nestedness of the model by checking that a lower bound on  $\nu$  satisfies a certain condition.

Fix  $y_0 < y_1$  (where  $y_0, y_1 \in Y$ ),  $k_0 \in D_y c(X, y_0)$  and set  $k_{max}(y_0, y_1, k_0) = \sup\{k : X_{\geq}(y_0, k_0) \subseteq X_{\geq}(y_1, k)\}$ . We then define the *minimal mass difference*,  $D_{\mu}^{min}$ , as follows:

$$D_{\mu}^{min}(y_0, y_1, k_0) = \mu(X_{\geq}(y_1, k_{max}(y_0, y_1, k_0)) \setminus X_{\geq}(y_0, k_0)).$$

The minimal mass difference represents the smallest amount of mass that can lie between  $y_0$  and  $y_1$ , and still have the corresponding level curves  $X_{=}(y_0, k_0)$  and  $X_{=}(y_1, k_1)$  not intersect.

**Theorem 1 (Theorem 3 [15]).** *Assume that  $\mu$  and  $\nu$  are absolutely continuous with respect to Lebesgue measure. If  $D_{\mu}^{min}(y_0, y_1, k(y_0)) < \nu([y_0, y_1])$  for all  $y_0 < y_1$  where  $k(y)$  is defined by (5), then  $(c, \mu, \nu)$  is nested. Conversely, if  $(c, \mu, \nu)$  is nested, we must have  $D_{\mu}^{min}(y_0, y_1, k(y_0)) \leq \nu([y_0, y_1])$  for all  $y_1 > y_0$ .*

then the following condition on the density of  $\nu$  follows

**Corollary 1 (Corollary 4 [15]).** *Assume that  $\mu$  and  $\nu$  are absolutely continuous with respect to Lebesgue measure. If for each  $y_0 \in Y$ , we have*

$$\sup_{y_1 \in Y, y_0 \leq y \leq y_1} \left[ \frac{D_\mu^{\min}(y_0, y_1, k(y_0))}{y_1 - y_0} - \nu(y) \right] < 0,$$

then  $(c, \mu, \nu)$  is nested.

### 3 A Variational Approach to Cournot-Nash Equilibria

We briefly recall here the variational approach to Cournot-Nash via Optimal Transport proposed in [4–6]. Agents are characterised by a type  $x$  belonging to a compact metric space  $X$  which is endowed with a given probability measure  $\mu \in \mathcal{P}(X)$  which gives the distribution of types in the agent population. For sake of simplicity we will consider probability measures absolutely continuous with respect to the Lebesgue measure and, with a slight abuse of notation we will identify the measure with its density. Each agent must choose a strategy  $y$  from a space strategy space  $Y$  (a compact metric space again) by minimizing a given cost  $\Phi : X \times Y \times \mathcal{P}(Y) \rightarrow \mathbb{R}$ . Notice that the cost  $\Phi(x, y, \nu)$  of one agent does not depend only on the type of the agent and the chosen strategy but also on the other agents' choice through the probability distribution  $\nu$  resulting from the whole population strategy choice. Then, an equilibrium can be described by a joint probability distribution  $\gamma$  on  $X \times Y$  which gives the joint distribution of types and strategies and is consistent with the cost-minimising behaviour of agents.

**Definition 2 (Cournot-Nash equilibrium).** Given a cost  $\Phi : X \times Y \times \mathcal{P}(Y) \rightarrow \mathbb{R}$ , a probability measure  $\mu \in \mathcal{P}(X)$ , a Cournot-Nash equilibrium is a probability measure on  $\mathcal{P}(X \times Y)$  such that

$$\gamma(\{(x, y) \in X \times Y : \Phi(x, y, \nu) = \max_{z \in Y} \Phi(x, z, \nu)\}) = 1, \tag{7}$$

and  $\pi_x(\gamma) = \mu, \pi_y(\gamma) = \nu$ .

Here we consider Cournot-Nash equilibria in the separable case that is

$$\Phi(x, y, \nu) = c(x, y) + F(\nu, y),$$

where  $c$  is a transport cost and the term  $F(\nu, y)$  captures all the strategic interactions. In particular  $F$  can be of the form

$$F(\nu, y) = f(\nu) + \int_Y \phi(y, z) d\nu(z),$$

where

- $f(\nu)$  captures the congestion effects: frequently played strategies are costly;
- $\int_Y \phi(y, z) d\nu(z)$  captures the positive interactions.

*Remark 2.* In order to understand better the role played by the term  $F(v, y)$ , let us consider again the example of the physicians and the cities that we explained in the introduction. In this case the distribution  $v$  of the types of the cities, namely the strategies the physician have to choose, is unknown. The main idea of the Cournot-Nash equilibria is to find the matching  $\gamma$  such that the agents minimize a transport cost (again this can represent the cost of commuting) and an extra term which takes into account

- a repulsive effect (the congestion): it is self-defeating opening a practice in a city chosen already by many other physicians;
- an attractive effect (the positive interactions term): the physicians want to share their own experiences with their peers.

In [6] the authors observed that  $F$  is the first variation of the energy

$$\mathcal{F}(v) = \int_Y f(v(y))dy + \int_{Y \times Y} \phi(y, z)dv(y)dv(z),$$

and consequently proved that the condition defining the equilibria is in fact the Euler-Lagrange equation for the following variational problem

$$\inf\{ \mathcal{J}(\mu, v) \mid v \in \mathcal{P}(Y) \}, \tag{8}$$

where

$$\mathcal{J}(\mu, v) = \mathcal{I}_c(\mu, v) + \mathcal{F}(v).$$

Indeed if  $v$  solves (8) and  $\gamma$  is the optimal transport plan between  $\mu$  and  $v$  for the cost  $c(x, y)$  then  $\gamma$  is a Cournot-Nash equilibrium for the cost  $\Phi(x, y, v)$  defined as above. Notice that once the cost function satisfies the twisted condition (also known in economics as the Spence-Mirrlees condition) then  $\gamma$  is determinist, that is  $\gamma$  is a pure equilibrium (see, for example, [18]).

### 3.1 Nestedness of Cournot-Nash Equilibrium

We now briefly recall the main results in [15] assuring that nestedness condition holds. Moreover, in the next section, we will see how nestedness can lead to the development of numerical methods to solve (8).

The congestion case

We firstly consider the case  $n = 1$ , meaning that we are in the multi-to-one Optimal Transport problem, and the functional  $\mathcal{F}(v)$  takes into account only the congestion effects. Indeed we consider the following functional

$$\mathcal{F}(v) = \int_Y f(v)dy$$

with  $f : [0, \infty) \rightarrow \mathbb{R}$  continuously differentiable on  $[0, \infty)$ , strictly convex with superlinear growth at infity and satisfying

$$\lim_{x \rightarrow 0^+} f'(x) = -\infty.$$

An example is the entropy  $f(v) = v \log(v)$ . In particular we have this result establishing lower and upper bounds on the density  $v$  and the nestedness of the model.

**Theorem 2 (Theorem 11 [15]).** Assume that  $v \in \mathcal{P}(Y)$ , with  $Y = (0, \bar{y})$  is a minimizer of (8) then  $v$  is absolutely continuous w.r.t. the Lebesgue measure and there exist two constants  $M_1$  and  $M_2$  depending on the cost function and  $Y$  such that

$$(f')^{-1}(M_1) \leq v(y) \leq (f')^{-1}(M_1)$$

. Moreover,  $(c, \mu, v)$  is nested provided

$$\sup_{y_1 \in Y, y_0 \leq y \leq y_1} \frac{D_{\mu}^{min}(y_0, y_1, k(y_0))}{y_1 - y_0} - (f')^{-1}(M_1) < 0$$

for all  $y_0 \in Y$ .

Notice that once we have established the model is nested then by setting  $k(y) = v'(y)$  we get

$$v(y) = \int_{X=(y, k(y))} \frac{D_{yy}^2 c(x, y) - k'(y)}{|D_{xy}^2 c(x, y)|} \mu(x) d\mathcal{H}^{m-1}(x) := G(y, k(y), k'(y)) \quad (9)$$

and differentiating the first order condition of (8) we get a second order differential equation for  $k$

$$k(y) + f''(G(y, k(y), k'(y))) \frac{d}{dy} G(y, k(y), k'(y)) = 0.$$

In Sect. 4.1 we detail how helpful nestedness is in order to design a suitable numerical method.

The interaction case

We consider now the case  $m > n \geq 1$  and a functional  $\mathcal{F}$  which captures only the interaction effects, that is

$$F(y, v) = V(y) + \int_Y \phi(y, z) dv(z),$$

where  $\phi$  is symmetric that is  $\phi(y, z) = \phi(z, y)$ . In this case it is actually more difficult to find lower and upper bounds as in the cogestion, but still the following nestedness result holds

**Theorem 3 (Theorem 15 [15]).** Assume that  $y \mapsto c(x, y) + V(y) + \phi(y, z)$  is uniform convex throughout  $X \times Y \times Y$  and

$$\langle (D_y c(x, y) + DV(y) + D_y \phi(y, z)) \mid n_Y(y) \rangle \geq 0 \quad \forall x \in X, z \in Y, y \in \partial Y,$$

where  $n_Y(y)$  is the outward normal to  $Y$ . Then the model  $(c, \mu, v)$  is nested.

Moreover in this special case one can characterize the minimizer by using the best-reply scheme that we will detail in Sect. 4.2.

## 4 Numerics

We now describe three different numerical approaches to solve (8). In particular we are now considering discretized measures that is  $\mu = \sum_{i=1}^M \mu_i \delta_{x_i}$  ( $x_i \in \mathbb{R}^m$ ) and  $\nu = \sum_{i=1}^N \nu_i \delta_{y_i}$  ( $y_i \in \mathbb{R}^n$ ), where  $M$  and  $N$  is the number of gridpoints chosen to discretize  $X$  and  $Y$  respectively. To reduce the amount of notation here, we use the same notations for the continuous problem as for the discretized one where integrals are replaced by finite sums and  $c$  and  $\gamma$  are now  $M \times N$  matrices.

### 4.1 Congestion Case

In this section we deal with a numerical method exploiting the nestedness result we provided for the congestion case. Consider the

$$\mathcal{F}(\nu) = \int_Y f(\nu) dy$$

with  $f$  satisfying the hypothesis we gave in the previous section and  $Y = (0, \bar{y})$  such that the model is nested. Then, the main idea of the iterative algorithm is actually to combine the optimality condition with the fact that in the nestedness case we can explicitly build the optimal transport map and the Kantorovich potential. We firstly recall that in this case the optimality condition of (8) reads

$$\nu(y) + f'(\nu) = C. \tag{10}$$

Since the model is nested we have that  $\nu(y) = \int_0^y k(s) ds$  which leads to

$$\nu(y) = (f')^{-1} \left( C - \int_0^y k(s) ds \right). \tag{11}$$

The idea is now to use (11) to obtain the following iterative algorithm: fix an initial density  $\nu^{(0)}$

$$\begin{cases} k^{(n)}(y) \text{ s.t } \mu(X_{\geq}(y, k^{(n)}(y))) = \nu^{(n-1)}((0, y]), \\ \nu^{(n)}(y) = (f')^{-1} (C^{(n)} - \int_0^y k^{(n)}(s) ds), \end{cases} \tag{12}$$

where  $C^{(n)}$  is computed such that  $\nu^{(n)}$  is a probability density.

*Remark 3.* This algorithm is actually an adaptation to the unequal dimensional case of the one proposed in [5]. For both the equal and unequal case we proved in [16] a convergence result by choosing a suitable metric.

*Remark 4.* Take  $f(\nu) = \nu \log(\nu)$  then the update of  $\nu^{(n)}$  reads

$$\nu^{(n)}(y) = \frac{\exp(-\int_0^y k^{(n)}(s) ds)}{\int_Y \exp(-\int_0^y k^{(n)}(s) ds) dy}$$

Numerical results

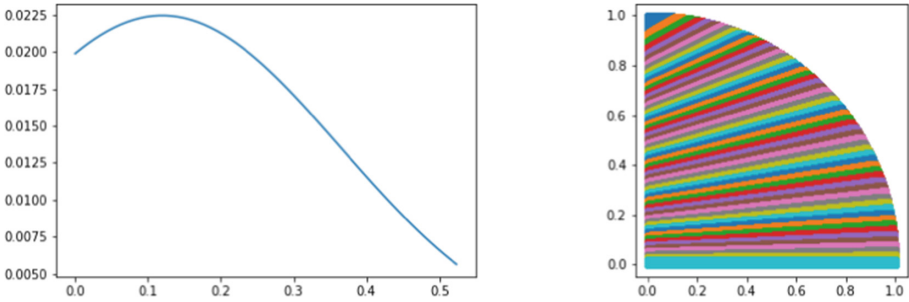
In all the simulations we present in this paragraph we have taken the uniform density on the arc  $(0, \frac{\pi}{6})$  as initial density  $v^{(0)}$ ,  $\mu$  is always the uniform on quarter disk  $X = \{x_1, x_2 > 0 : x_1^2 + x_2^2 < 1\}$  and  $f(v)$  is the entropy as in the remark above. By Corollary 12 in [15], when the cost is given by  $c(x, y) = |x_1 - \cos(y)|^2 + |x_2 - \sin(y)|^2$  we know that the model is nested provided  $\bar{y} \leq 0.61$  (which is exactly the case we consider for the numerical tests). Moreover, we have also considered an additional potential term, which will favour a concentration in a certain area of  $(0, \frac{\pi}{6})$ , so that the functional  $\mathcal{F}$  has the form

$$\mathcal{F}(v) = \int_Y f(v)dy + \int V(y)d\nu(y),$$

where  $V(y) = 10|y - 0.1|^2$ . The second equation in (12) then becomes

$$v^{(n)}(y) = (f')^{-1}\left(C^{(n)} - V(y) - \int_0^y k^{(n)}(s)ds\right).$$

In Fig. 1 we show the final density we have obtained (on the left) and the intersection between the level-set  $X_{\geq}(y, k^*(y))$  for the final solution and the support of the fix measure  $\mu$ . Notice that, as expected, the level set  $X_{\geq}(y, k^*(y))$  do not cross each other meaning that the model is nested.



**Fig. 1.** (Left) Final density  $v^*$ . (Right) Intersection between the level-set  $X_{\ge}(y, k^*(y))$  for the final solution and the support of the fix measure  $\mu$

In Fig. 2, we present some simulations by taking the same data set as above, but the cost function is now given by

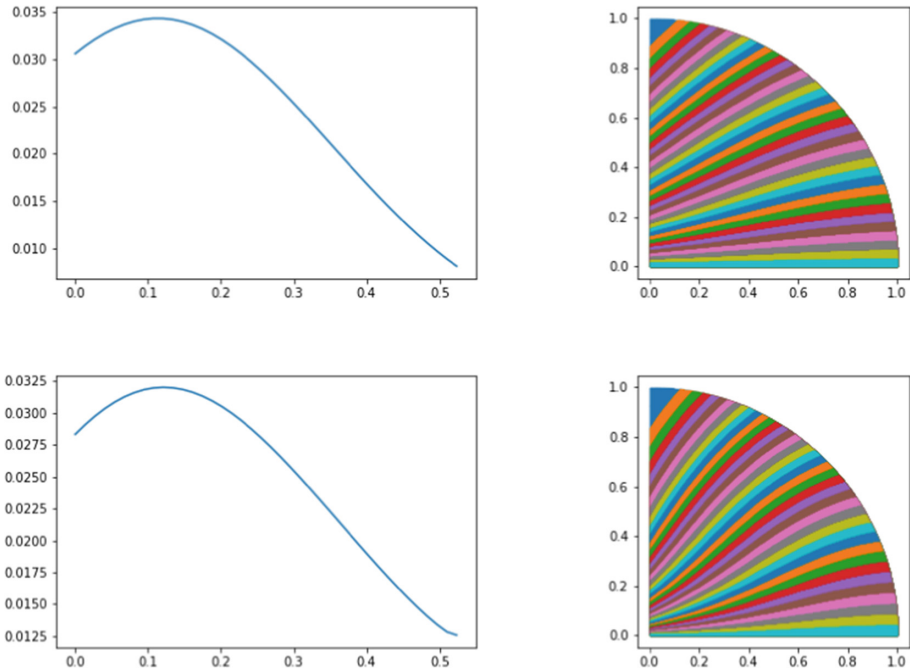
$$c(x, y) = \left(1 + |x_1 - \cos(y)|^2 + |x_2 - \sin(y)|^2\right)^{\frac{p}{2}}$$

with  $p > 2$ . Notice that now the level set are not straight lines anymore.

**4.2 Best Reply Scheme**

Assume now that

$$F(y, v) = V(y) + \int_Y \phi(y, z)d\nu(z)$$



**Fig. 2.** First row:  $p = 4$ .(Left) Final density  $v^*$ . (Right) Intersection between the level-set  $X_{\geq}(y, k^*(y))$  for the final solution and the support of the fix measure  $\mu$ . Second row:  $p = 8$ .(Left) Final density  $v^*$ . (Right) Intersection between the level-set  $X_{\geq}(y, k^*(y))$  for the final solution and the support of the fixed measure  $\mu$ .

and consider the first optimality condition of (8) which reads

$$c(x, y) + F(y, v) = C, \tag{13}$$

where  $C$  is a constant assuring that the minimizer  $v$  is a probability distribution. Notice that by differentiating (13) with respect to  $y$  we get

$$D_y c(x, y) + DF(y, v) = 0, \tag{14}$$

we denote by  $B_v : X \rightarrow Y$  the map such that

$$D_y c(x, B_v(x)) + DF(v, B_v(x)) = 0, \tag{15}$$

which is well defined under the hypothesis of Theorem 3. Then, the best-reply scheme, firstly introduced in [5], consists in iterating the application defined as

$$\mathcal{B}(v) := (B_v)_\# \mu. \tag{16}$$

We refer the reader to [15][Theorem 19] for a convergence result. Notice that [15][Theorem 19] assures also that  $\mathcal{B}(v)$  is absolutely continuous with respect to the

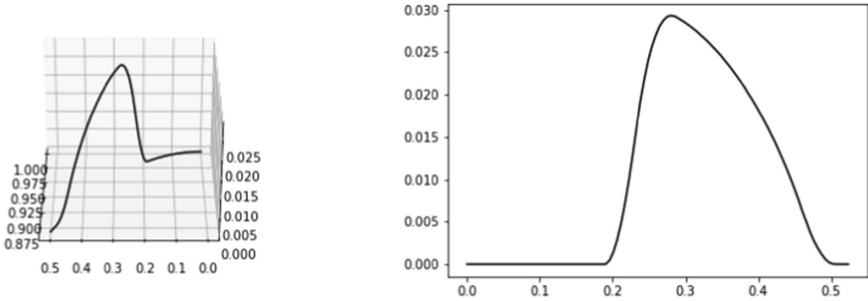


Lebesgue measure, meaning that the density, and so (16), can be computed by using the co-area formula. We obtain then the following algorithm: fix  $v^{(0)}$  then

$$v^{(n+1)} = \mathcal{B}(v^{(n)}) := (B_{v^{(n)}})_{\#}\mu.$$

**Numerical results**

As in the previous section the initial density  $v^{(0)}$  is the uniform on the arc  $(0, \frac{\pi}{6})$  and  $\mu$  is the uniform on  $X = \{x_1, x_2 > 0 : x_1^2 + x_2^2 < 1\}$ . Moreover, we take here a potential  $V(y) = |y - \frac{\pi}{12}|^2$  and a quadratic interaction  $\phi(y, z) = |y - z|^2$ . In Fig. 3 we plot the final density obtained in the case in which the cost function is  $c(x, y) = |x_1 - \cos(y)|^2 + |x_2 - \sin(y)|^2$ . Notice that since there is not a congestion term like the entropy, the final density has a support much smaller than the interval  $(0, \frac{\pi}{6})$ .



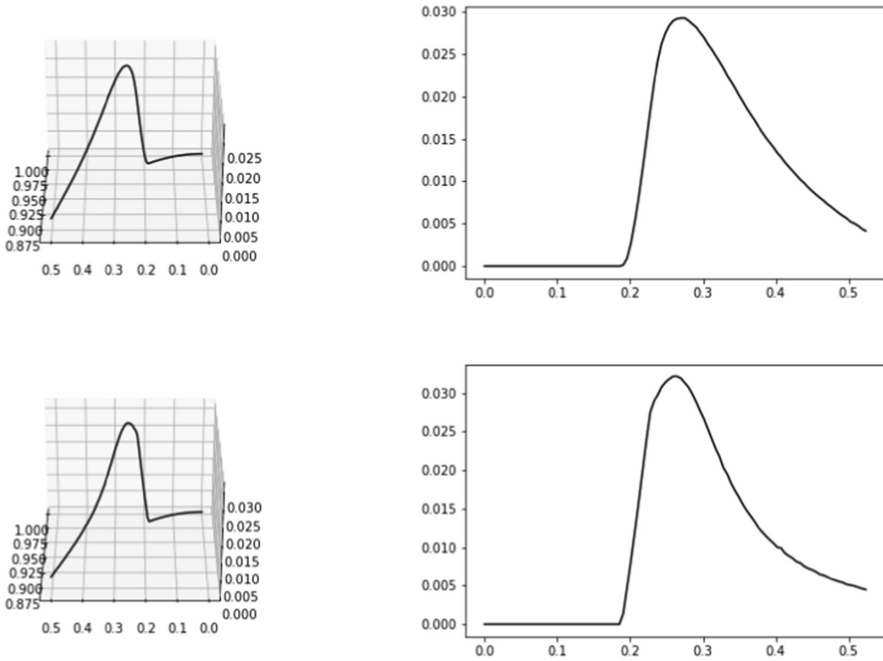
**Fig. 3.** (Left) Final density  $v^*$  concentrated on the arc  $(0, \frac{\pi}{6})$ . (Right) Final density  $v^*$ .

In Fig. 4 we repeat the simulation with the same data set but taking the cost function  $c(x, y) = \left(1 + |x_1 - \cos(y)|^2 + |x_2 - \sin(y)|^2\right)^{\frac{p}{2}}$  with  $p = 4, 8$ . Notice that the cost function plays the role of an attractive term with respect to the fix measure  $\mu$  which makes the final density have a more spread support.

**4.3 The Entropic Regularization**

The last numerical approach we consider is the entropic regularization of Optimal Transport (we refer the reader to [3, 6, 10, 11, 17] for more details) which can be easily used to solve (8) no matter the choice of the transportation cost  $c(x, y)$  and the functional  $\mathcal{F}(v)$ . This actually implies that the entropic regularization is clearly more flexible than the methods we have previously introduced but in some cases (i.e. the congestion term is the entropy) the major drawback of this approach is an extra diffusion. The entropic regularization of Optimal Transport can be stated as follows

$$\mathcal{I}_{c,\varepsilon}(\mu, \nu) := \inf \left\{ \int_{X \times Y} c d\gamma + \varepsilon \int_{X \times Y} (\log(\gamma) - 1) d\gamma \mid \gamma \in \Pi(\mu, \nu) \right\}, \quad (17)$$



**Fig. 4.** (Left) Final density  $\nu^*$  concentrated on the arc  $(0, \frac{\pi}{6})$ . (Right) Final density  $\nu^*$ .

where we assume that  $0(\log(0) - 1) = 0$ . The problem now is strictly convex and it can be re-written as

$$\mathcal{I}_{c,\varepsilon}(\mu, \nu) := \inf \{ \varepsilon \mathcal{H}(\gamma | \bar{\gamma}) \mid \gamma \in \Pi(\mu, \nu) \},$$

where  $\mathcal{H}(\gamma | \bar{\gamma}) := \int_{X \times Y} (\log(\frac{\gamma}{\bar{\gamma}}) - 1) d\gamma$  is the relative entropy and  $\bar{\gamma} = \exp(\frac{-c}{\varepsilon})$ . It can be proved that (17) is strictly convex problem and it admits a solution of the form

$$\gamma^* = \exp\left(\frac{u^*}{\varepsilon}\right) \bar{\gamma} \exp\left(\frac{v^*}{\varepsilon}\right),$$

where  $u^*$  and  $v^*$  are the Lagrange multipliers associated to the marginal constraints. We also highlight that by adding the entropic term we have penalized the non-negative constraint meaning that the solution  $\gamma^*$  is always positive.

The regularized version of (8) can be stated in the following way

$$\inf \{ \varepsilon \mathcal{H}(\gamma | \bar{\gamma}) + \mathcal{F}(\pi_y(\gamma)) + \mathcal{G}(\pi_x(\gamma)) \} \tag{18}$$

where  $\mathcal{F}$  is the functional capturing the congestion and interactions effects and  $\mathcal{G}(\rho) = i_\mu(\rho)$  is the indicator function in the convex analysis sense and it is used to enforce the prescribed first marginal. Before introducing the generalization of Sinkhorn algorithm proposed in [10] we briefly recall without proof a classical duality result.

**Proposition 1.** *The dual problem of (18)*

$$\sup_{u,v} -\mathcal{G}^*(-u) - \mathcal{F}^*(-v) - \varepsilon \int_{X \times Y} \exp\left(\frac{u^*}{\varepsilon}\right) \bar{\gamma} \exp\left(\frac{v^*}{\varepsilon}\right) dx dy. \tag{19}$$

Moreover strong duality holds.

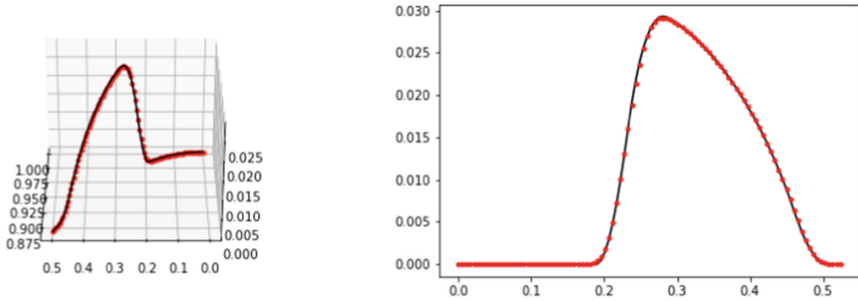
The generalized Sinkhorn algorithm is then obtained by relaxation of the maximizations on the dual problem (19). We get the iterative method computing a sequence of potentials: given 2 vectors  $u^{(0)}$  and  $v^{(0)}$ , then the update at step  $n$  is defined as

$$\begin{cases} u^{(n)} := \operatorname{argmax}_u -\mathcal{G}^*(-u) - \varepsilon \int_{X \times Y} \exp\left(\frac{u}{\varepsilon}\right) \bar{\gamma} \exp\left(\frac{v^{(n-1)}}{\varepsilon}\right), \\ v^{(n)} := \operatorname{argmax}_v -\mathcal{F}^*(-v) - \varepsilon \int_{X \times Y} \exp\left(\frac{u^{(n)}}{\varepsilon}\right) \bar{\gamma} \exp\left(\frac{v}{\varepsilon}\right). \end{cases} \tag{20}$$

*Remark 5.* For many interesting functionals  $\mathcal{F}$  the relaxed maximizations can be computed point-wise in space and analytically. Notice that a problem can arise in treating the interaction term, but one can easily overcome this difficulty by adopting a semi-implicit approach as the one proposed in [7].

**Numerical results**

$\mu$  is the uniform measure on  $X = \{x_1, x_2 > 0 : x_1^2 + x_2^2 < 1\}$ . Moreover, we take here a functional  $\mathcal{F}$ , as in the best reply scheme section, given by the sum of a potential  $V(y) = |y - \frac{\pi}{12}|^2$  and a quadratic interaction  $\phi(y, z) = |y - z|^2$ . We can then compare the numerical results obtained by the best reply scheme and generalized Sinkhorn, with a regularization parameter  $\varepsilon = 10^{-3}$ , in the case of the quadratic cost function (Fig. 5).



**Fig. 5.** (Left) Final density  $v^*$  obtained by the best reply scheme (dark solid line) and by entropic regularization (red dots) concentrated on the arc  $(0, \frac{\pi}{6})$ . (Right) Final density  $v^*$  obtained by the best reply scheme (dark solid line) and by entropic regularization (red dots).

*Remark 6.* Notice that we have not compared the iterative method for the congestion case with the entropic regularization. This is almost due to the fact that adding an other entropy term to the problem will cause an extra diffusion which could be difficult to treat even by choosing a small regularization parameter.

**Acknowledgements.** B.P. is pleased to acknowledge the support of National Sciences and Engineering Research Council of Canada Discovery Grant number 04658-2018. L.N. was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH and from the CNRS PEPS JCJC (2022).

## References

1. Aumann, R.: Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* **32**, 39–50 (1964)
2. Aumann, R.: Markets with a continuum of traders. *Econometrica* **34**, 1–17 (1966)
3. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.* **37**(2), A1111–A1138 (2015)
4. Blanchet, A., Carlier, G.: From nash to cournot-nash equilibria via the monge-kantorovich problem. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **372**(2028), 20130398 (2014)
5. Blanchet, A., Carlier, G.: Remarks on existence and uniqueness of Cournot-Nash equilibria in the non-potential case. *Math. Financ. Econ.* **8**(4), 417–433 (2014)
6. Blanchet, A., Carlier, G.: Optimal transport and Cournot-Nash equilibria. *Math. Oper. Res.* **41**(1), 125–145 (2016)
7. Blanchet, A., Carlier, G., Nenna, L.: Computation of cournot-nash equilibria by entropic regularization. *Vietnam J. Math.* **46**, 15–31 (2017)
8. Chiappori, P.A., McCann, R., Pass, B.: Multidimensional matching. ArXiv e-prints (2016)
9. Chiappori, P.A., McCann, R.J., Pass, B.: Multi-to one-dimensional optimal transport. *Commun. Pure Appl. Math*
10. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Scaling algorithms for unbalanced transport problems. Technical report (2016). <http://arxiv.org/abs/1607.05816>
11. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst. (NIPS)* **26**, 2292–2300 (2013)
12. Galichon, A., Salanié, B.: Matching with trade-offs: Revealed preferences over competing characteristics. Technical report (2009). Preprint SSRN-1487307
13. Mas-Colell, A.: On a theorem of Schmeidler. *J. Math. Econ.* **3**, 201–206 (1984)
14. McCann, R.J., Pass, B.: Optimal transportation between unequal dimensions. arXiv preprint [arXiv:1805.11187](https://arxiv.org/abs/1805.11187) (2018)
15. Nenna, L., Pass, B.: Variational problems involving unequal dimensional optimal transport. *J. Math. Pures Appl.* **139**, 83–108 (2020)
16. Nenna, L., Pass, B.: Transport type metrics on the space of probability measures involving singular base measures. arXiv preprint [arXiv:2201.00875](https://arxiv.org/abs/2201.00875) (2022)
17. Peyré, G.: Entropic approximation of Wasserstein gradient flows. *SIAM J. Imaging Sci.* **8**(4), 2323–2351 (2015)
18. Santambrogio, F.: Optimal transport for applied mathematicians. *Calculus of variations, PDEs, and modeling*. In: *Progress in Nonlinear Differential Equations and their Applications*, vol. 87. Springer, Cham (2015). <https://doi.org/10.1007/978-3-319-20828-2>
19. Villani, C.: Optimal Transport: Old and New. *Grundlehren der mathematischen Wissenschaften*, vol. 338. Springer, Cham (2009). <https://doi.org/10.1007/978-3-319-20828-2>



# Stacking Regression for Time-Series, with an Application to Forecasting Quarterly US GDP Growth

Erkal Ersoy, Haoyang Li, Mark E. Schaffer<sup>(✉)</sup>, and Tibor Szendrei

Edinburgh Business School, Heriot-Watt University, Edinburgh, UK  
m.e.schaffer@hw.ac.uk

**Abstract.** Machine learning methods are being increasingly adopted in economic forecasting. Many learners are available, and a practical issue now presents itself: which one(s) to use? The answer we suggest is ‘stacking regression’ (Wolpert, 1992), an ensemble method for combining predictions of different learners. We show how to use stacking regression in the time series setting. Macroeconomic and financial time series data present their own challenges to forecasting (extreme values, regime changes, etc.), and this presents challenges to stacking as well. Our findings suggest that using absolute deviations for scoring the base learners performs well compared to stacking on mean squared error. We illustrate this with a Monte Carlo exercise and an empirical application: forecasting US GDP growth around the Covid-19 pandemic.

**Keywords:** Stacking regression · machine learning · forecasting · robust statistics

## 1 Introduction

Machine learning methods are being imported in applied econometrics in a variety of settings. These methods provide powerful tools for prediction and forecasting. This poses a new problem for applied econometricians: too much choice. There are many machine learning estimators available. Which learner should they use? Model selection methods typically select a model and then conduct inference based on the assumption that the model actually generated the data. Their inference can only be trusted if the ‘best’ model selected happens to be a close approximation to the true data generating process. In practice, however, it is more likely that the best model captures some aspects of the truth, while other models capture other aspects. By conditioning only on the best model, model selection methods ignore all the evidence contained in the alternatives and can lead to misleading results in the sense of being either systematically

---

Invited paper for the Sixth International Econometrics Conference of Vietnam, Banking University of Ho Chi Minh City, Vietnam, 9–11 January 2023. All errors are our own.

wrong or overfitting the data. Whenever quantities that are not model-specific are of interest, therefore, it makes more sense to create a mixture of the different models rather than select a ‘best’ model (Steel, 2020). To this end we consider ‘stacking regression’ (Wolpert, 1992), an ensemble method for combining predictions of different learners, as a way to mix information contained in the different models. Stacking regression is, in effect, a generalization of cross-validation.

The dependent data setting has its own peculiarities: financial and macroeconomic data are typically serially correlated, extreme values are an issue, etc. Tuning methods have to avoid data leakage (letting information from the future leak into the model training process). This applies to stacking regression on two fronts: (1) the training of the different learners needs to avoid data leakage from the future into the individual learners; (2) the stacking procedure that combines the predictions of these learners has to avoid data leakage. We outline in this paper how this is done in practice.

Lastly, this paper examines alternative scoring approaches, i.e., how in practice weights are assigned to the different learners to obtain the stacked forecast. The most common scoring approach in both cross-validation and stacking is mean-squared prediction error (MSPE). In the time-series setting, however, the mean absolute prediction error (MAPE) is an attractive alternative. Macro and financial time series commonly present researchers with problems such as extreme values and regime changes, and working with absolute rather than squared deviations has been shown in other contexts to add robustness to the analysis. Stacking on mean squared has been used in some empirical applications (Pavlyshenko, 2018; Ribeiro et al., 2019; Ribeiro and dos Santos Coelho, 2020; da Silva et al., 2020), but we are unaware of any systematic exploration of stacking on median. We examine whether the robustness of the median extends to stacking regression. Using Monte Carlo experiments and an empirical application focused on US GDP around the COVID-19 pandemic, we find that stacking on the regression has attractive features for practitioners.

The paper is organised as follows. We first briefly summarize how cross-validation is done in the time-series setting, and then introduce stacking regression. In our applications we use 5 different ‘base learners’ that are combined to obtain a stacking forecast, and the next section describes them in brief: lasso, ridge, elastic net, support vector machine, and random forest. The next section presents the results of a Monte Carlo exercise that shows that stacking regression using MAPE for scoring compares favourably to stacking regression when MSPE is used for evaluation. We then present the results of a practical application – forecasting US GDP growth – and again show that stacking regression using MAPE for scoring performs well. The final section concludes.

## 2 Time Series Cross-Validation

*Cross-validation* (CV) allows researchers to choose a model specification – typically, a tuning parameter (e.g., the penalization parameter for lasso) – based on predictive performance, and it is often used to avoid modelling difficulties

like overfitting and selection bias. As part of the process, the dataset is split into two sections: the ‘training’ sample, which is used to fit the model, and the ‘validation’ or ‘holdout’ sample, which is used to assess predictive performance. Mean squared prediction error, MSPE, is a common choice of metric for this.

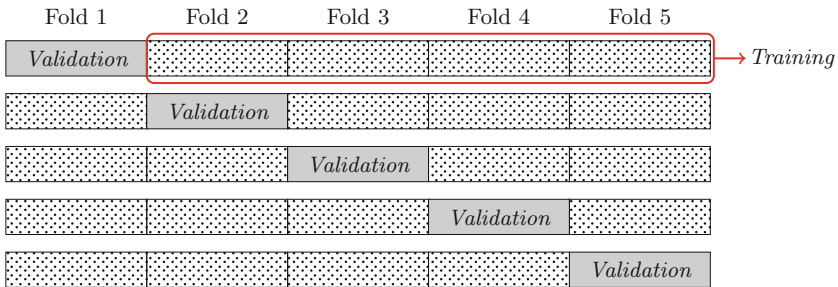
This resampling technique iteratively trains and tests a model using different portions of the data to tune the parameter of the base learner. The goal is to find the model that has the best out-of-sample predictive performance and can generalize to other samples from the same population. Since stacking can be regarded as a logical extension of cross-validation, we briefly go through CV before formally introducing the implementation of stacking.

In the case of independent data, ‘ $K$ -fold’ cross-validation is the most commonly used approach. In  $K$ -fold cross-validation, the data are randomly split into  $K$  portions or ‘folds’. At each iteration, one fold is treated as the validation set while the remaining  $K - 1$  folds are treated as the training set to fit the model for some value of the tuning parameter. After each fold is used as the validation set once (and only once), the predictive performance of the model is estimated by averaging the MSPE over all the validation sets. Given a range of values for the tuning parameters, the model with the best predictive performance is selected as the final model (Fig. 1).

The iterative nature of cross-validation makes it computationally intensive: the model needs to be repeatedly estimated and its performance checked across different folds and across a grid of values for the tuning parameter.

Cross-validation with dependent data, i.e., in a time series setting, adds further complications because of the need to ensure that the validation data are independent of the training data. The key issue with  $K$ -fold cross-validation in the context of time series prediction is data leakage. When data are dependent, the information from the validation set can leak into the training set, leading to overfitting and hence poor generalization.

With the exception of very specific cases where  $K$ -fold cross-validation may be appropriate, researchers should typically use time series cross-validation, a version of ‘non-dependent cross validation’ (Bergmeir et al., 2018) where cross-validation is set up to account for the nature of the dependence that may see



**Fig. 1.**  $K$ -fold cross-validation for cross sectional data. ( $K = 5$ )

dependent observations omitted from the validation sets. In simple terms, time series cross-validation ensures that the training and validation take place with  $h$ -step-ahead forecasts. For example, 1-step-ahead cross-validation (Hyndman and Athanasopoulos, 2018) fits a model on  $t$  observations and assesses predictive performance based on the forecast for time  $t + 1$ .

More generally, consider a series of validation sets, each of which includes one observation at  $t + 1$ . The corresponding training set of each validation set would then consist of observations through time  $t$ , all of which, by definition, will have occurred before  $t + 1$ . After the predictions at current iteration are made, the validation set moves forward by one and the current observation is added to the training set to form the new training set for the next iteration. Thus, future observations are never used to forecast previous ones; data never leaks into the training process from the future. After going through all the predetermined validation sets, the model with the best predictive performance is selected as the final model. This one-step-ahead expanding window approach is demonstrated in Fig. 2(a).

This process can be generalized for  $h$ -step-ahead CV and the training window can be fixed instead of expanding. A rolling window fixes the size of the training set by deleting the most distant observation when a new observation is added, while an expanding window simply adds the new observation to the current training set. Therefore, the rolling window always has a fixed training size predetermined by the researcher and the expanding window includes a growing number of observations in the training set. The rolling window is useful when the series is volatile or the forecasting depends largely on the most recent history, while the expanding window is more appropriate when the series has a stable trend or seasonal pattern.

Figure 2 and Fig. 3 show examples of one-step- and two-step-ahead forecasts with expanding and fixed windows, e.g., Fig. 3(b) demonstrates 2-step-ahead CV with a fixed window, where ‘ $T$ ’ and ‘ $V$ ’ refer to the training and validation samples, respectively. The first step is identical across Fig. 2 and Fig. 3: observations 1 to 3 constitute the training set and observation 4 is used for validation while the remaining observations are unused as indicated by a dot (‘.’). In step 2, the training set becomes larger in the expanding window setup such that it consists of observations 1 through 4 (Fig. 2(a)) whereas the size of the training set is fixed in Fig. 3(a) such that it consists of observations 2 through 5. Considering the focus here is on macroeconomic data, and GDP in particular, we opt for an expanding window approach for the base learners in this paper. For the Monte Carlo experiments we will use 1-step ahead forecasts, while for the empirical application we will use 1 and 4 step ahead forecasts.



		Step				
		1	2	3	4	5
1		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
2		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
3		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
<i>t</i>	4	<i>V</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
	5	.	<i>V</i>	<i>T</i>	<i>T</i>	<i>T</i>
	6	.	.	<i>V</i>	<i>T</i>	<i>T</i>
	7	.	.	.	<i>V</i>	<i>T</i>
	8	.	.	.	.	<i>V</i>

(a)  $h = 1$ , expanding window

		Step				
		1	2	3	4	5
1		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
2		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
3		<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
<i>t</i>	4	.	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
	5	<i>V</i>	.	<i>T</i>	<i>T</i>	<i>T</i>
	6	.	<i>V</i>	.	<i>T</i>	<i>T</i>
	7	.	.	<i>V</i>	.	<i>T</i>
	8	.	.	.	<i>V</i>	.
	9	.	.	.	.	<i>V</i>

(b)  $h = 2$ , expanding window

**Fig. 2.** Rolling  $h$ -step ahead cross-validation with expanding training window. ‘ $T$ ’ and ‘ $V$ ’ denote that the observation is included in the training and validation sample, respectively. A dot (‘.’) indicates that an observation is excluded from both training and validation data.

		Step				
		1	2	3	4	5
1		<i>T</i>	.	.	.	.
2		<i>T</i>	<i>T</i>	.	.	.
3		<i>T</i>	<i>T</i>	<i>T</i>	.	.
<i>t</i>	4	<i>V</i>	<i>T</i>	<i>T</i>	<i>T</i>	.
	5	.	<i>V</i>	<i>T</i>	<i>T</i>	<i>T</i>
	6	.	.	<i>V</i>	<i>T</i>	<i>T</i>
	7	.	.	.	<i>V</i>	<i>T</i>
	8	.	.	.	.	<i>V</i>

(a)  $h = 1$ , fixed window

		Step				
		1	2	3	4	5
1		<i>T</i>	.	.	.	.
2		<i>T</i>	<i>T</i>	.	.	.
3		<i>T</i>	<i>T</i>	<i>T</i>	.	.
<i>t</i>	4	.	<i>T</i>	<i>T</i>	<i>T</i>	.
	5	<i>V</i>	.	<i>T</i>	<i>T</i>	<i>T</i>
	6	.	<i>V</i>	.	<i>T</i>	<i>T</i>
	7	.	.	<i>V</i>	.	<i>T</i>
	8	.	.	.	<i>V</i>	.
	9	.	.	.	.	<i>V</i>

(b)  $h = 2$ , fixed window

**Fig. 3.** Rolling  $h$ -step ahead cross-validation with fixed training window.

### 3 Stacking

Machine learning methods have become popular in time series applications, too. With ‘wide’ databases becoming available for different applications, there has been an influx of applied papers utilising a plethora of different methods for fit and variable selection (Goulet Coulombe et al., 2022; Kohns and Bhattacharjee, 2022; Massacci and Kapetanios, 2023). Given the breadth of choice, one question to ask is which model to use. In some cases we have prior information which can

help us make a decision, e.g., we believe the problem at hand is sparse and linear, and so we prefer the lasso. But often we don't have this information. Stacking regression (Wolpert, 1992) provides a potential solution: rather than select a 'best' model, mix the different models in a principled manner. In essence, the idea is that our models describe reality given some simplification. Not all models simplify the problem at hand along the same 'axis' (i.e., the models have different assumptions). In such cases, we can find an optimal convex mixture of the models to give an overall better prediction. Importantly, since stacking is a generalization of cross-validation, this convex combination of models is theoretically founded (for a textbook treatment, see Hastie et al. (2009)).

With stacking regression, we combine predictions from multiple learners into a meta model. The initial set of models consists of 'base learners' or 'Level 0 models'. The stacking method combines the predictions of the base learners into a 'meta model' or 'Level 1 model'. Assume we have  $M$  base learners and denote by  $\hat{f}_m(\mathbf{x}_i)$  the prediction for observation  $i$  of base learner  $m$  after tuning. Formally, stacking can be represented as:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \min_{w_1, \dots, w_M} \sum_{i=1}^n \left( y_i - \sum_{m=1}^M w_m \hat{f}_m(\mathbf{x}_i) \right)^2 \\ \text{s.t. } w_m &\geq 0, \\ \sum_m |w_m| &= 1 \end{aligned} \tag{1}$$

Note that this set of equations essentially describes a constrained least squares problem, where we constrain the weights to be non-negative and their sum to be unity. These constraints lead to better performance and facilitate the interpretation of stacking as a weighted average of base learners (Hastie et al., 2009).

Stacking on mean squared is the most common approach here, just as it is the most common choice of scoring method in standard cross-validation. However, time series applications are notorious for various breaks occurring in the data. Extreme events are known to have a large influence on models focused on the mean (Rousseeuw and Hubert, 2011). Importantly, from a forecasting perspective, conditional mean models usually yield subpar forecasts right after crisis episodes, which is exactly the moment policymakers and stakeholders need accurate information. The median has a better breakdown point than the mean (Huber and Ronchetti, 2009; Rousseeuw and Hubert, 2011), which makes it a more attractive choice for modelling during uncertain times. Recasting the equation to stack on the median yields:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \min_{w_1, \dots, w_M} \sum_{i=1}^n \left| y_i - \sum_{m=1}^M w_m \hat{f}_m(\mathbf{x}_i) \right| \\ \text{s.t. } w_m &\geq 0, \\ \sum_m |w_m| &= 1 \end{aligned} \tag{2}$$

Note how the only difference between Eq. (1) and (2) is that the objective function in the latter minimises absolute deviations. From an application standpoint this can be solved using Koenker and Ng (2005) instead of a constrained least squares. To avoid confusion with learners and scoring methods, we refer to stacking on the median as “Stacking (L1)” and stacking on mean squared as “Stacking (L2)” in the figures and tables.

## 4 Base Learners

In this section we briefly describe the base learners we use in the paper: lasso, ridge, elastic net, support vector machine, and random forest.

### 4.1 Lasso

When dealing with high-dimensional data, researchers often have to circumvent overfitting problem. Including too many irrelevant variables in the regression model can result in poor out of sample generalization. Tibshirani (1996) introduced the lasso (‘least absolute shrinkage and selection operator’) to improve the prediction accuracy and interpretability of models in such case by performing both regularization and variable selection. Consider a sample with  $n$  observations and  $p$  covariates. Let  $y_i$  be the outcome and  $X_i$  be the vector of regressors for the  $i^{\text{th}}$  observation. The lasso solves the optimization problem:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - X_i' \beta)^2 + \lambda \|\beta\|_1 \right\}, \quad (3)$$

where  $\beta$  is the coefficient vector, and  $\lambda$  is the regularization parameter that controls the overall penalty level. A higher  $\lambda$  means a stronger penalty on the magnitude of all coefficients. At one extreme, lasso estimates approach those of OLS as  $\lambda$  goes to zero. At the other end, all coefficients are shrunk to zero when  $\lambda$  is large enough. In practice,  $\lambda$  is usually tuned through cross-validation, while the search range needs to be predetermined by the researcher. Including non-zero coefficients for covariates will increase the score of the loss function. Consequently, all coefficients will shrink towards zero, while the coefficients of those covariates who contribute little or nothing to the outcome will be shrunk to exactly zero. The covariates are typically standardized so that the solution does not depend on the measurement scale. The  $L_1$  norm  $\|\beta\|_1$  makes lasso a quadratic programming problem, hence there is no closed form solution and the computation can be slow. Lasso is an appropriate choice for both prediction and variable selection when the model is sparse.

### 4.2 Ridge

Ridge regression (Tikhonov, 1963; Hoerl and Kennard, 1970) was the most popular technique for improving predictive performance prior to lasso. It resembles

lasso but has one key difference: the  $L_2$  norm is used for the penalization term. In particular, ridge solves the problem:

$$\hat{\boldsymbol{\beta}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - X'_i \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \quad (4)$$

Ridge also intends to reduce prediction error by shrinking all coefficients towards zero, but no coefficients will be shrunk to exactly zero, which means that it does not perform variable selection. Additionally, no requirement for the sparsity assumption makes ridge attractive when the model is dense. Unlike the lasso, ridge is also computationally efficient since it has a closed form solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y},$$

where  $\mathbf{X}$  is the  $n \times p$  design matrix,  $\mathbf{I}$  is a  $p \times p$  identity matrix, and  $\mathbf{y}$  is the  $n \times 1$  vector of outcome. In general, the solution is still well defined when  $\mathbf{X}'\mathbf{X}$  is rank deficient provided that  $\lambda$  is sufficiently large.

### 4.3 Elastic Net

Elastic net regularization (Zou and Hastie, 2005) is simply a linear combination of  $L_1$  and  $L_2$  penalties of lasso and ridge. Specifically, it solves the problem:

$$\hat{\boldsymbol{\beta}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - X'_i \boldsymbol{\beta})^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\} \quad (5)$$

Note that lasso has several limitations in practice. Firstly, for high-dimensional data where the number of covariates  $p$  is larger than the number of observations  $n$ , lasso selects at most  $n$  covariates before it saturates. Secondly, lasso tends to select only one covariate from a group of highly correlated covariates and discards the others even they all contribute to the outcome. Thirdly, the solutions of lasso are not always unique and re-ordering the covariates may end up with different estimates. The elastic net overcomes the limitations by adding a quadratic term to the penalty while still preserving the advantages of lasso. The  $L_2$  penalty makes the loss function above strongly convex and hence has a unique solution. A common practice of reparameterization is to set:

$$\begin{aligned} \lambda_1 &= \alpha \lambda \\ \lambda_2 &= (1 - \alpha) \lambda \end{aligned} \quad (6)$$

where  $\lambda$  controls the overall penalty level and  $\alpha$  controls the balance between lasso and ridge. A higher  $\alpha$  indicates a higher weight on lasso and more coefficients will be shrunk to zero. The reparameterization is useful in the sense that it allows us to fix  $\alpha$  and select a single parameter  $\lambda$  instead of tuning  $\lambda_1$  and  $\lambda_2$  separately.

#### 4.4 Support Vector Machine

The linear *Support Vector Machine (SVM)* (Boser et al., 1992) solves a classification problem by finding a decision function,  $f(X)$ , based on a set of  $n$  observations  $X_i$  with labels  $y_i \in \{1, -1\}$ , that divides all observations into two classes. From the training set this algorithm estimates the parameters of the decision function  $f(X)$  through a learning process. Then the classification of a new observation is predicted according to the decision function. Each data point  $X_i$  can be viewed as a  $p$ -dimensional vector. Consider a  $p - 1$  dimensional *hyperplane* defined by  $\{X : f(X) = X'\beta + \beta_0 = 0\}$ , linear SVM chooses the best hyperplane that maximizes the distance (or margin) from it to the nearest data point in each class. The hyperplane is a geometric representation of the decision function  $f(X)$  with a  $p$ -dimensional norm vector,  $\beta$ , and a bias term,  $\beta_0 \in \mathbb{R}$ . Linear SVM's training outcome is a classification rule,  $G(X)$ , depending on the side of the hyperplane that an unclassified observation lands on. In particular,  $G(X) = \text{sign}[f(X)] = \text{sign}[X'\beta + \beta_0]$ .

If there exists two parallel hyperplanes that separate the two classes of training set, linear SVM maximises the *margin*,  $M = \frac{2}{\|\beta\|}$ , between the planes by solving the minimization problem:

$$\begin{aligned} & \underset{\beta, \beta_0}{\text{minimize}} \quad \|\beta\|_2^2 \\ & \text{s.t. } y_i(X_i'\beta + \beta_0) \geq 1 \quad \forall i \in \{1, 2, \dots, n\} \end{aligned} \quad (7)$$

However, data are sometimes not linearly separable. This can be incorporated in the optimization function by including a hinge loss function  $\xi_i = \max(0, 1 - y_i(X_i'\beta + \beta_0))$ , which is proportional to the distance from the margin if the point is misclassified and takes the value of zero if the point lies on the correct side. We can now modify the optimization problem above as:

$$\begin{aligned} & \underset{\beta, \beta_0, \xi}{\text{minimize}} \quad \|\beta\|_2^2 + C \sum_i^n \xi_i \\ & \text{s.t. } y_i(X_i'\beta + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \in \{1, 2, \dots, n\}, \end{aligned} \quad (8)$$

where  $C$  is the *cost parameter* that penalizes the amount of observations inside the margin. A larger value of  $C$  will make the optimization choose a smaller *margin* and hence increasing the overfitting. The Lagrange dual function can be written as:

$$\begin{aligned} \mathcal{L}(\alpha_1, \alpha_2, \dots, \alpha_n) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i (X_i' X_j) y_j \alpha_j \\ & \text{s.t. } \alpha_i \geq 0, \quad \sum_i^n \alpha_i y_i = 0 \quad \forall i \in \{1, 2, \dots, n\} \end{aligned} \quad (9)$$

where  $\alpha_i$  are Lagrange multipliers. The function can be efficiently solved by quadratic programming algorithms, yielding  $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i y_i X_i$ . Notably,  $\alpha_i \geq 0$

only if the point  $X_i$  lies on the boundary of the *margin*. such points are called *support vectors*.

A regression version of SVM, *support vector regression (SVR)*, was proposed by Drucker et al. (1996). Training SVR means solving the problem:

$$\begin{aligned} & \underset{\beta, \beta_0}{\text{minimize}} \quad \frac{1}{2} \|\beta\|_2^2 \\ & \text{s.t.} \quad |y_i - X_i' \beta - \beta_0| \leq \epsilon \quad \forall i \in \{1, 2, \dots, n\}, \end{aligned} \quad (10)$$

where  $\epsilon$  is a tuneable parameter that serves as a threshold. The distance between any prediction and the true value should be within the range  $\epsilon$ .

## 4.5 Random Forests

The *random forests* (Breiman, 2001) algorithm applies ‘bagging’ (bootstrap aggregation) to *decision tree learners*. Although tree learners are invariant to transformations of features and hence robust to inclusion of irrelevant features, they tend to overfit the training sets and suffer from high variance. In principle, tree-based algorithms split the training set into subsets based on thresholds of selected features with the purpose to minimize the prediction error. The process is recursively repeated on each derived subset until a subset (node) with the minimum amount of observations is reached, which is usually set to five for regression problem. However, the minimum size can be tuned to alleviate overfitting.

Given a training set  $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$  with outcomes  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ , random forests repeatedly draw a random sample of size  $n$  with replacement from the set  $B$  times and fits trees to the samples. A tree learner  $T_b$  is trained on each sample  $\mathbf{Z}_b = (\mathbf{y}_b, \mathbf{X}_b)$ . After training, the prediction for an observation with features  $X$  can be made by averaging over all the individual trees:

$$\hat{T}^B(X) = \frac{1}{B} \sum_{b=1}^B T_b(X) \quad (11)$$

The bootstrapping procedure improves the model performance in the sense that it reduces the variance without increasing the bias. The predictions of a single tree could be sensitive to outliers, but the average of trees will be less affected, as long as trees are uncorrelated. Although bootstrap sampling decorrelates the trees by training them on different samples, they can still be highly correlated if several strong features are mostly or always selected. Random forests addresses this problem by including another type of bagging: a modified tree learning algorithm that selects a random subset of the features at each split is used. The process is called *feature bagging*. Typically, the number of randomly selected features is set to  $m = \frac{p}{3}$  for regression, where  $p$  is the total number of features. Additionally, the optimal number of trees  $B$  can be tuned by finding the one that minimizes the *out-of-bag error (OOB)*, which is the mean prediction

error on each observation  $X_i$ , using only the trees that do not include  $X_i$  in their bootstrap sample.

In this paper, rather than explicitly tune the random forest specifications by selecting the number of features or tree depth using cross-validation, we specify a small number of different random forest specifications as base learners. In this sense, the stacking algorithm tunes the random forest specification as part of the overall stacking procedure when it assigns weights to the different random forest learners.

## 5 Monte Carlo

We investigate the finite-sample performance of stacking using simulated data. Stacking is compared with its components: lasso, ridge, elastic net, support vector machine, and random forest. The Monte Carlo designs are explained in Sect. 5.1, the normalization of data is presented in Sect. 5.2, and evaluation criterion are discussed in Sect. 5.3. Finally, we compare the results of different learners in Sect. 5.4.

### 5.1 Setup

The dependent variable  $y_t$  is generated as:

$$y_t = \alpha y_{t-1} + \beta X_{t-1} + \epsilon_t, \text{ for } t = 1, 2, \dots, n, \quad (12)$$

where  $X_{t-1}$  is the vector of all variables of length  $p$  and  $\epsilon_t$  is the error term.  $X_t$  follows multivariate normal distribution  $N_p(0, \Sigma)$ , where  $\Sigma$  is the  $p \times p$  covariance matrix with element  $\Sigma_{ij} = 0.2^{|i-j|}$ . Considering that the performances of learners such as lasso can be largely different based on whether the model is sparse or dense, we set  $\beta = \{1, 0.5, 0.2, 0, 0, 0, \dots\}$  to simulate the sparse model and  $\beta = \{\frac{1}{\sqrt{1}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \dots\}$  to represent the dense model. The following three types of data generating processes are used:

**DGP I (Autoregressive distributed lag model):**  $\alpha$  is fixed to be 0.5. The error term is distributed as  $\epsilon_t \sim N(0, 1)$ .

**DGP II (ARDL model with a fat tail):** Since extreme events are likely to happen in many applications, it is of interest to know how the performances change with more outliers. Now the error term follows a  $t$  distribution with degrees of freedom set to 3.

**DGP III (Non-stationary model):** The dependent variable  $y_t$  is generated as in DGP I, but now  $\alpha$  is set to be 1 such that the model is non-stationary.

We run  $R = 100$  simulations for each DGP with  $p = n = 200$ . Since a ‘bad’ starting point may over-sample points occurring with low probability before it reaches the equilibrium distribution, we also include 50 burn-in periods at the beginning of each DGP. 30% of the data (60 observations) are treated as validation sets, on which stacking regression is done. For each simulation, the

last observation is treated as the testing set, i.e. this observation is not used for training the base learners, nor for the stacking regression. Instead, the test set is used to evaluate the performance of the different estimators.

## 5.2 Data Normalization

Normalization is important for distance-based machine learning algorithms such as SVM. A distance summarizes the relative difference between two vectors. Numerical values may have different scales, which can greatly affect the calculation of distance measures. In particular, features with relatively larger scales will have stronger impacts on the distance even they actually contribute less to the dependent variable. To avoid this issue, all the features are normalized to have mean zero and unit variance. Note that scaling the whole sample up front will lead to data leakage when the data are dependent. Therefore, normalization needs to be conducted for each training set individually within the corresponding cross-validation split. Once the stacking regressors are trained, we then normalize the whole training sample at once, and the corresponding normalization factors are applied to normalize the hold-out sample. In essence, we run all our base learners on the normalized data without causing any data leakage issue.

## 5.3 Performance Measures

Root mean squared prediction error (RMSPE) is used to compare the finite-sample performance of two types of stacking and its base learners. Specifically,

$$RMSPE = \sqrt{\frac{\sum_{t=1}^T (y_t - \hat{y}_t)^2}{T}},$$

$$MAPE = \frac{\sum_{t=1}^T |y_t - \hat{y}_t|}{T},$$

where  $T$  is the number of predictions, i.e. the size of the testing set.  $y_t$  and  $\hat{y}_t$  are the out-of-sample realized and predicted values respectively.<sup>1</sup>

## 5.4 Results

Table 1 shows the results from MC experiments. Overall, both fat tails and non-stationarity will lead to higher RMSPE of all the learners. Lasso always performs the best among all the base learners when the model is sparse. While ridge or elastic net has the smallest mean RMSPE when the model is dense. As expected, both stacking methods follow closely the best base learners in different situations, and sometimes even outperform all the base learners. Even though the difference

<sup>1</sup> Note that  $T = 1$  in our MC setting, leading to  $RMSPE = MAPE$ . As such we will only focus on RMSPE when discussing the MC results. We define both measures here because in the empirical application,  $T$  equals to 30, leading to  $RMSPE \neq MAPE$ .



**Table 1.** mean RMSPE of Monte Carlo Experiments

	Sparse			Dense		
	DGP I	DGP II	DGP III	DGP I	DGP II	DGP III
lasso	1.543	2.765	2.434	3.453	9.748	18.218
ridge	2.494	4.155	11.347	2.561	10.155	29.241
EN	1.550	2.778	2.575	3.435	9.700	18.197
svm	2.598	4.446	14.268	3.842	11.806	36.933
rf10	2.398	3.666	6.261	8.335	12.809	22.939
rf50	2.246	3.173	5.675	7.681	13.448	21.803
rf100	2.248	3.281	5.485	7.430	12.998	20.761
rf200	2.219	3.351	5.301	7.342	13.221	21.300
Stacking (L1)	1.611	2.686	2.487	2.808	10.078	17.064
Stacking (L2)	1.611	2.693	2.469	2.822	9.873	17.284

is small, stacking on median seems to perform slightly better than stacking on mean in general. It is worth noting that stacking on median has a smaller mean RMSPE than stacking on mean in DGP II where the model is sparse and has a fat tail, since the median method is more robust to outliers. However, the contrary is true when the model is dense, suggesting that the degree of sparsity may also influence the relative performance of two stacking methods. A further investigation is beyond the scope of the present paper.

## 6 Empirical Application

For this empirical application we use McCracken and Ng (2020), a database of US macroeconomic variables at the quarterly frequency. Kohns and Szendrei (2020) have shown that one can use the median as an adequate measure of fit on this database. Here, we take the mantle forward with the performance of the models when stacking on the median and compare the performance with stacking on the mean square. Our application includes the Covid-19 period, which is notoriously difficult to incorporate in forecasting models (Primiceri and Tambalotti, 2020; Ioannidis et al., 2022). In this section, we use the same base learners described in the Monte Carlo section above.

To ensure a rich selection of variables, we follow Kohns and Szendrei (2020) and start our empirical exercise from 1970Q1, which means we have 228 variables. We use the final 30 observations for testing purposes. Importantly, this means that the Covid-19 period is included in the testing set as well as data from ‘normal’ times. We perform 1 quarter ahead and 1 year (4 quarters) ahead direct forecasts.

The results for the static forecast exercises are presented in Fig. 4 for the 1-quarter-ahead, and Fig. 5 for the 1-year-ahead forecast horizon. The figures show the root mean squared prediction error (RMSPE) and the mean absolute

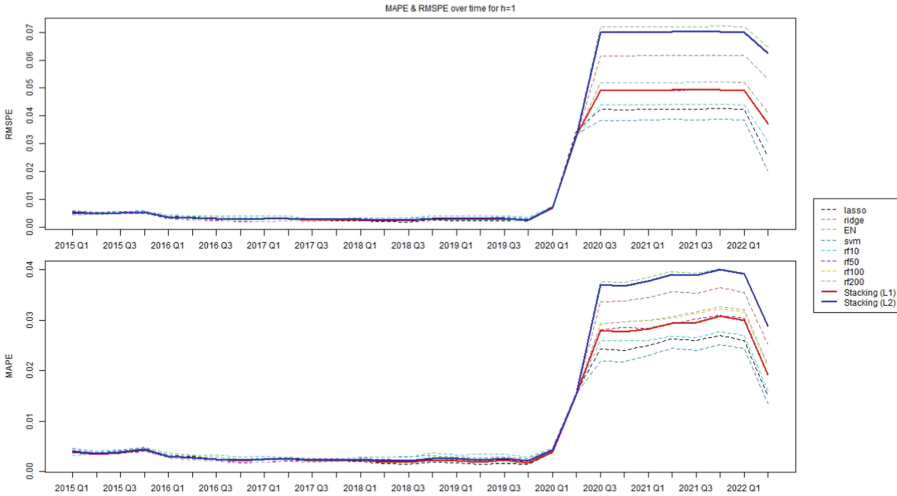


Fig. 4. Forecast results for 1 quarter ahead

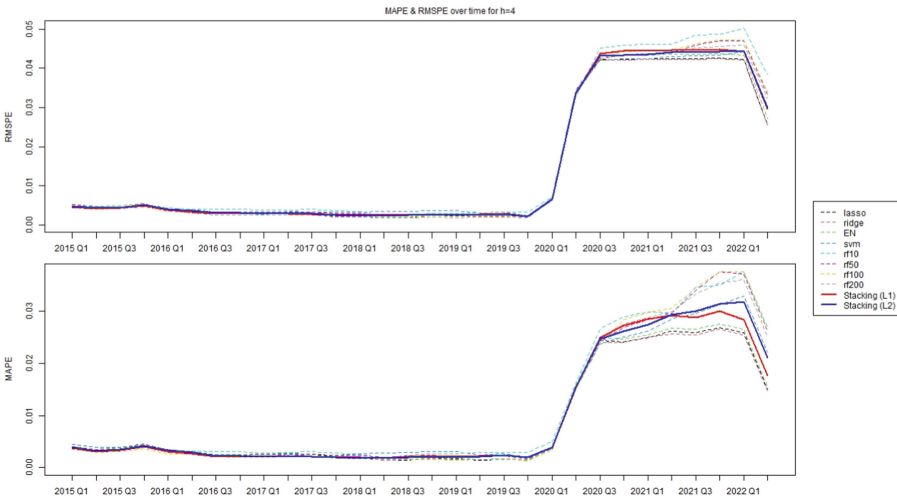
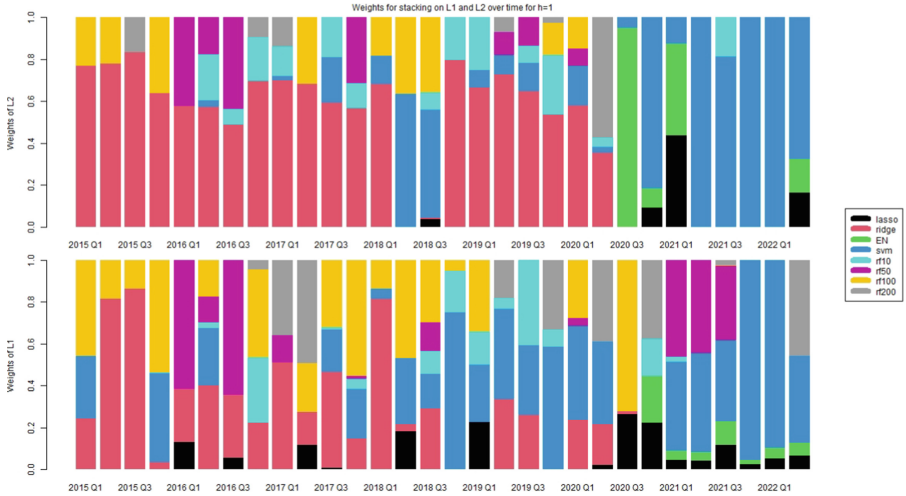


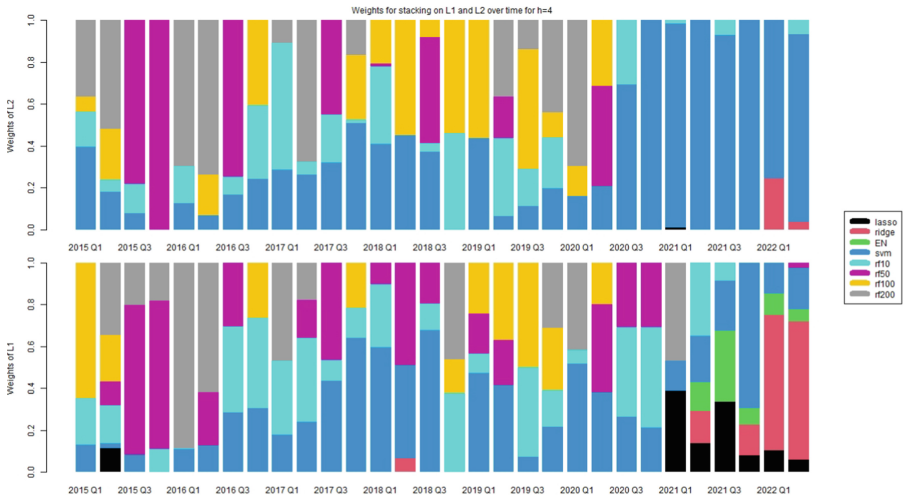
Fig. 5. Forecast results for 1 year ahead

prediction error (MAPE) with a moving window of 2 years. Note that our evaluation window is always backwards looking. The figures reveal that the base learners that work particularly well in ‘normal’ times (e.g., elastic net and ridge regression), show worse performance during the Covid-19 period, while some models perform better around periods of crisis (e.g., lasso). This highlights that choosing a model that performs adequately across all time periods is difficult.

Looking at the stacked regression performance in the short forecast horizon, we can see that during normal times stacking on the mean and median



**Fig. 6.** Weights of the different stacking methods (1 quarter ahead)



**Fig. 7.** Weights of the different stacking methods (1 year ahead)

offers comparable performance. Nevertheless, during the crisis episode, the performance of the two stacking methods deviates. During the Covid-19, stacking on the median offers far better performance than stacking on the mean. This is not surprising given the fact that the median is more robust to rank preserving shocks (Huber and Ronchetti, 2009). The results are similar for the longer forecast horizon: stacking on the median performs admirably.

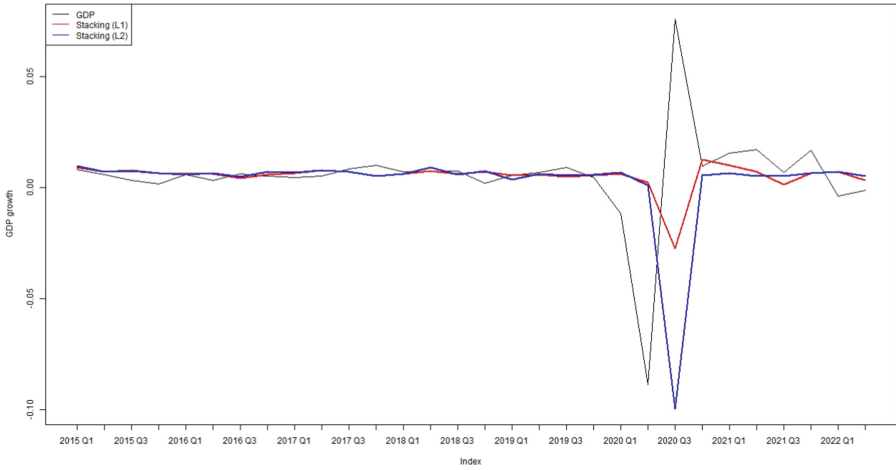
Interestingly, the performance of stacking on the median is not impacted by the choice of the horizon: RMSPE and MAPE are almost identical throughout the time-frame. Looking at the weights in Figs. 6 and 7 reveals why this might be the case. In essence, stacking on the median is more likely to mix information from more base learners. Given that longer forecast horizons have more uncertainty associated with it, a stacking method that is more likely to incorporate information from a more diverse set of models is likely to fare better.

Comparing the weights across the forecast horizons also reveals how different types of base learners are preferred at the different forecast horizons, with the SVM being the only base learner that is included frequently for both horizons. At the shorter forecast horizon, ridge regression is more dominant especially for stacking on mean square, while for the longer forecast horizons random forests are far more prevalent. Random forests being selected at the longer forecast horizon is likely because although random forests overfit in-sample, this has little to no consequences out-of-sample.<sup>2</sup> Importantly, we can see from the weights that although there are multiple random forests among the base learners, stacking regression always assigned non-zero weights to other base learners as well. This further highlights a key advantage of stacking: we are not limited to mixing only one type of model. In this instance, information from the elastic net learner is mixed into the stacking regression, which leads to better performance, especially during the crisis episodes.

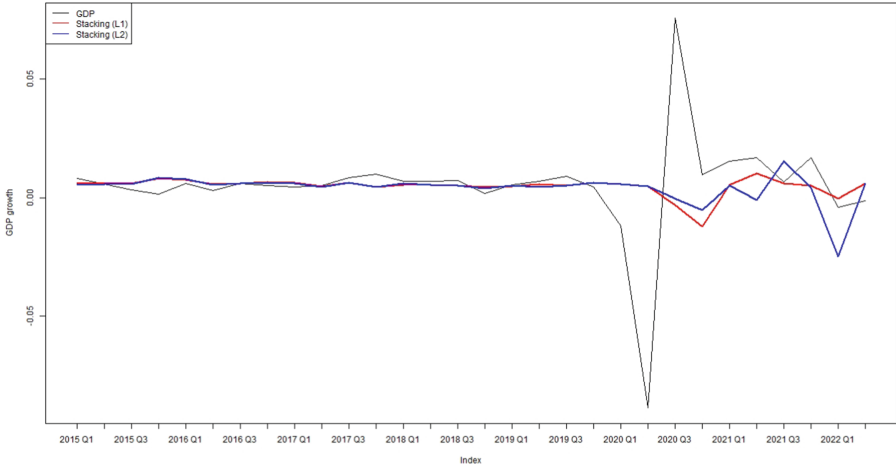
Our results point towards running both types of stacking methods at all times and relying on stacking during crisis times due to its robustness. It is difficult to know *ex ante* (and sometimes even in real time) whether one is in a crisis, which makes it difficult to choose between the two stacking methods. Figures 8 and 9 show that the fitted values of the two stacking methods are very close during normal times, but deviate from each other during crisis periods. The tendency for the two sets of fitted values to deviate during crises episodes is not too surprising given the robustness of the median to outliers (Huber and Ronchetti, 2009). As such, one can opt to favor stacking on the median, when the fitted values deviated from each other. We leave for future research the question of at what point deviations between the fitted values should be considered significant from a policy maker perspective.

---

<sup>2</sup> See Goulet Coulombe (2020) for further discussion and an explanation of why random forests tend to perform relatively well in a forecasting setting.



**Fig. 8.** Fitted values and GDP at the 1 quarter ahead horizon



**Fig. 9.** Fitted values and GDP at the 1 year ahead horizon

## 7 Conclusion

The key goal of stacking regression is to obtain an optimal mix of models that can lead to better fit than one particular model. Stacking regression has been popular for cross-section data and in this paper we outlined how to apply the method to time-series data. We note that extreme observations are not infrequent in time-series settings (e.g., macroeconomics and finance), and this can have detrimental effects on models focused on optimizing the squared residual. To remedy this we propose ‘stacking on the median’, since the median is more robust to outliers Rousseeuw and Hubert (2011).

In the Monte Carlo exercise we find that stacking on the median performs admirably, even beating stacking on the mean squared error. These results are corroborated by the empirical application focused on US GDP forecasts around the global pandemic. Rather than exclusively stacking on the median, we propose that policymakers consider running the two methods simultaneously, as the fitted values deviate during uncertain times. This way the policymaker will have information on not just what the forecasted value is, but also when an extreme event has occurred.

## References

- Bergmeir, C., Hyndman, R.J., Koo, B.: A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **120**, 70–83 (2018)
- Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152 (1992)
- Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
- Da Silva, R.G., Ribeiro, M.H.D.M., Fraccanabbia, N., Mariani, V.C., dos Santos Coelho, L.: Multi-step ahead bitcoin price forecasting based on vmd and ensemble learning methods. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE (2020)
- Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **9** (1996)
- Goulet Coulombe, P.: To bag is to prune (2020). arXiv preprint [arXiv:2008.07063](https://arxiv.org/abs/2008.07063)
- Goulet Coulombe, P., Leroux, M., Stevanovic, D., Surprenant, S.: How is machine learning useful for macroeconomic forecasting? *J. Appl. Econom.* **37**(5), 920–964 (2022)
- Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. SSS, Springer, New York (2009). <https://doi.org/10.1007/978-0-387-84858-7>
- Hoerl, A.E., Kennard, R.W.: Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**(1), 69–82 (1970)
- Huber, P.J., Ronchetti, E.M.: *Robust statistics*. Wiley Series in Probability and Mathematical Statistics (2009)
- Hyndman, R.J., Athanasopoulos, G.: *Forecasting: Principles and Practice* (2 ed.) (2018)
- Ioannidis, J.P., Cripps, S., Tanner, M.A.: Forecasting for COVID-19 has failed. *Int. J. Forecast.* **38**(2), 423–438 (2022)
- Koenker, R., Ng, P.: Inequality constrained quantile regression. *Sankhyā Indian J. Stat.* 418–440 (2005)
- Kohns, D., Bhattacharjee, A.: Nowcasting growth using google trends data: a Bayesian structural time series model. *Int. J. Forecast.* (2022)
- Kohns, D., Szendrei, T.: Horseshoe prior Bayesian quantile regression (2020). arXiv preprint [arXiv:2006.07655](https://arxiv.org/abs/2006.07655)
- Massacci, D., Kapetanios, G.: Forecasting in factor augmented regressions under structural change. *Int. J. Forecast.* (2023)
- McCracken, M., Ng, S.: FRED-QD: a quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research (2020)

- Pavlyshenko, B.: Using stacking approaches for machine learning models. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 255–258. IEEE (2018)
- Primiceri, G.E., Tambalotti, A.: Macroeconomic forecasting in the time of COVID-19. Manuscript, Northwestern University, pp. 1–23 (2020)
- Ribeiro, M.H.D.M., dos Santos Coelho, L.: Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Appl. Soft Comput.* **86**, 105837 (2020)
- Ribeiro, M.H.D.M., Ribeiro, V.H.A., Reynoso-Meza, G., dos Santos Coelho, L.: Multi-objective ensemble model for short-term price forecasting in corn price time series. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
- Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**(1), 73–79 (2011)
- Steel, M.F.: Model averaging and its use in economics. *J. Econ. Lit.* **58**(3), 644–719 (2020)
- Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996)
- Tikhonov, A.N.: On the solution of ill-posed problems and the method of regularization. In: *Doklady Akademii Nauk*, vol. 151, pp. 501–504. Russian Academy of Sciences (1963)
- Wolpert, D.H.: Stacked generalization. *Neural Netw.* **5**(2), 241–259 (1992)
- Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**(2), 301–320 (2005)



# Maximum Entropy Learning with Neural Networks

Woraphon Yamaka<sup>(✉)</sup>

Center of Excellence in Econometrics, Faculty of Economics,  
Chiang Mai University, Chiang Mai 50200, Thailand

**Abstract.** Conventionally, the back-propagation (BP), maximum likelihood (ML) and Bayesian approaches have been applied to train Artificial Neural Networks (ANN). This study presents a Generalized Maximum Entropy (GME) learning algorithm for ANN, designed specifically to handle limited training data and unknown error distribution. Maximizing only the entropy of parameters in the ANN allows more effective generalization capability, less bias towards data distributions, and robustness to over-fitting compared to the conventional algorithm learning. In the implementations, GME is compared with the conventional algorithms in terms of their forecasting performances in both simulation and real data studies. The findings demonstrate that GME outperforms other competing estimators when training data is limited and the distribution of the error is unknown.

**Keywords:** Artificial neural network · Comparison of estimators · Entropy

## 1 Introduction

Neural networks have received considerable attention in recent years, for being a self-learning and self-adaptive model with the powerful abilities in handling the nonlinear problem and complex issue (Chen et al. 2018; Ramos et al. 2021). Recently, the technique has been utilized in many purposes like prediction and classification (Chen et al. 2018; Ramos et al. 2021; Yamaka, Phadkantha, and Maneejuk 2021). In this study, I aim at introducing an alternative algorithm, which is the generalized maximum entropy estimation (GME) (Golan, Judge, and Miller 1996), to artificial neural networks (ANN) to improve the prediction performance.

Estimation of the neural network parameters is quite challenging as it needs to adjust the weights and biases to ensure that the output is close to the desired output (Lin et al. 2016). Many estimation techniques and concepts have been proposed and developed to tune weight and bias parameters in the neural networks (Chon and Cohen 1997). It should be noted that these parameters are both learnable parameters which are used to link the input layer, the hidden layer and the output together. For example, if we have a single layer network, the input data is multiplied with the weight parameter; then a bias is added before passing the transformed input data to the next hidden layer. Next, the output layer can be obtained by multiplying the transformed input data with another

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*,  
SSDC 483, pp. 150–162, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_8](https://doi.org/10.1007/978-3-031-35763-3_8)



weight parameter followed by the inclusion of an additional bias to obtain the output. Traditionally, parameters are estimated using the methods of back-propagation (BP) (White 1989), maximum likelihood (ML) (Gish 2020) and Bayesian (Müller and Insua 1998).

From the computational point of view, the BP algorithm minimizes the cost function, which is commonly assumed to be mean square error, in order to obtain the optimal parameters. Many iterative learning steps are required in the learning process to obtain a better learning performance. However, it is well known that given the cost function as mean square error, it leads to the strong assumption that all the feature components are equivalent (Wang, Du and Wang 2017). Thus, Gish (2020) proposed a probabilistic view of neural networks to derive the maximum likelihood estimation. Specifically, the cost function of the BP algorithm is replaced by the likelihood function. The basic concept of the ML method is that the optimal parameters should be chosen such that the probability of the observed sample data is maximized. This estimation has several attractive properties including: consistency, asymptotic normality, and efficiency when the sample size approaches infinity (Chen et al. 2013). Lin et al. (2016) argued that although the learning process of these estimators are generalized correctly to the new inputs after sufficient training, the learning speed is slow and is not incremental in nature (old input should still be trained with the new input) (Fu, Hsu, Principe 1996). Also, if we limit the training data to reduce the computational cost of the estimations and gain a better control over the training data, we may face the overfitting problem (Chu et al. 2021). It should be noted that overfitting occurs when the network has memorized the training input, but it has not learned to generalize to new inputs, leading to overconfident predictions. In the Bayesian approach, these issues can be handled in a natural and consistent way. The non-informative priors are used to handle the complexity of the data and network; as a result, the model is weighted by the posterior probability given the data sample. However, this estimation still suffers from some complicated problems such as the training time, the efficient parameter estimation, the random walk in the high-dimensional parameter cases (Kocadağlı 2015).

According to the above view about the estimation methods, despite these estimations generally perform well, they have inherent additional limitations (Kocadağlı and Aşıkil 2014; Lin et al. 2016; and Yang, Baraldi, and Zio 2016). First of all, in the cases of ML and Bayesian, determining the most suitable distribution (likelihood and posterior distributions) requires an expert, otherwise it is possible to construct the incorrect functional structure. Secondly, when the neural network model is being trained using the BP and ML, a large training data is required. Thirdly, it has often been found that BP and ML are prone to overfitting (Dorling et al. 2003). Thus, we need to limit the complexity of the network making it suitable to the learning problem defined by the data (Bishop 1995).

To overcome these limitations, GME is suggested to estimate the weight and bias parameters of ANN. GME-ANN can be one of the popular neural networks models for dealing with prediction problem. This study aims at investigating the possibility of developing a ANN model based on the use of GME. Unlike ML and Bayesian, before ANNs are being trained, the prior information regarding the likelihood and posterior distributions are not required. GME allows us to produce methods that are capable of learning

complex behaviors without human intervention. It also has an ability to fit the data without making specific assumptions; therefore, I hypothesized that estimation with GME (GMS-ANN) would enable the resulting parameter estimates to be more unbiased to data distributions and robust to over-fitting issues compared to those ML, and Bayesian. In addition, there are many pieces of evidence confirming the high estimation performance of GME (Alibrandi and Mosalam 2018; Maneejuk, Yamaka, and Sriboonchitta 2020), despite small sample size and limited training data. In this study, thus, the performance of each estimation approach and their relative performance with a focus on small sample sizes are investigated.

The rest of this paper is organized as follows. Section 2 describes the proposed methodology. Section 3 presents the experiment studies. The real data example is reported in Sect. 4. Finally, Sect. 5 provides the conclusion of this study.

## 2 Model Setup

The idea is to build an entropy function with a neural network constraint to replace the loss function or probability function discussed in the previous section. In other words, the GME is used as the estimator to adjust the weights and biases of the neural network by maximizing the Shannon entropy with the ANN equation constraint. In particular, weights and biases in ANN are reparametrized as the discrete random variables on bounded supports. The sum of entropy distributions of the weights and biases is maximized subject to model consistency constraints. The weights and biases of interest are then calculated as the expectation of random variables on the prescribed supports under the derived distributions of the entropy maximization.

### 2.1 ANN with Three Layers

In this section, I provide three layered ANN consisting of an input layer with  $I$  input neurons, one hidden layer with  $H$  hidden neurons, and one output layer, as the example. Mathematically, the hidden and input layers of ANN can be expressed as

$$y_i = \sum_{h=1}^H \left\{ \omega_h^O f^I \left( \sum_{k=1}^K \omega_{k,h}^I x_{i,k} + b_h^I \right) + b_h^O \right\} + \varepsilon_i, \quad (1)$$

where  $y_i$ , for  $t = 1, \dots, T$ , and  $x_{i,k}$ , for  $k = 1, \dots, K$ , are output and input variables, respectively.  $\omega_{k,h}^I$  is the weight parameter of input  $x_{i,k}$  that connects the input  $x_{i,k}$  and the  $h$ th neuron in the hidden layer,  $b_h^I$  is the bias for  $h$ th neuron in the hidden layer.  $f^I$  is the activation function that provides the nonlinearity to the ANN structure, and scales its received inputs to its output range. In this study, the logistic function is employed as it is easy to calculate and its first derivative is simple (Kocadağlı and Aşıkıl 2014). Likewise, I use  $\omega_h^O$  and  $b_h^O$  to denote weight and bias terms, respectively.  $\varepsilon_i$  is the error term.

Learning occurs through the adjustment of the path weights and node biases. Traditionally, all the weight and bias parameters are estimated by the BP method. The optimal parameters are estimated by minimizing the squared difference between observed output and estimated output. The loss function can be written as follows,

$$Loss = \frac{1}{N} \sum_{i=1}^N \left\{ y_i - \sum_{h=1}^H \left\{ \omega_h^{O,fI} \left( \sum_{k=1}^K \omega_{k,h}^I x_{i,k} + b_h^I \right) + b_h^O \right\} \right\}, \quad (2)$$

## 2.2 Maximum Entropy Learning for ANN Model

In this study, the maximum entropy (ME) of Jaynes (1982) is generalized to estimate weights and biases in the ANN equation. As I mentioned before, all parameters are calculated as the expectation of random variables on the prescribed supports under the derived distributions of the entropy. More precisely, the random variables are treated as the probabilities and the information entropy of these probabilities can be measured by Shannon's entropy (Shannon 1948)

$$H(\mathbf{p}) = - \sum_d p_d \log p_d, \quad (3)$$

where  $p_d$  is the probability of the possible outcome  $d$ , such that  $\sum_d p_d = 1$ . Under this maximum entropy principle, the distribution is chosen for which the information is just sufficient to determine the probability assignment. In addition, it seeks information within the data without imposing arbitrary restrictions. In this study, I follow the idea of Golan, Judge, and Miller (1996) and generalize the ME solution to the inverse problems with error, expressed in the ANN framework.

To estimate the unknown parameters in Eq. (1), say  $\omega_h^O$ ,  $\omega_{k,h}^I$ ,  $b_h^O$  and  $b_h^I$ , for  $h = 1, \dots, H$  and  $k = 1, \dots, K$ , we reparameterize them as the expectation of weights on the prescribed supports. The weight parameters. Each weight has a bounded support space,  $\mathbf{z}_{hk} = [z_{hk,1}, \dots, z_{hk,M}]$ , associated with the  $h$ th neuron and  $k$ th variable, which is symmetrically built around zero and weighted by the vector  $\mathbf{p}_{hk} = [p_{hk,1}, \dots, p_{hk,m}]$ . Note that  $z_{hk,1}$  and  $z_{hk,M}$  are, respectively, the lower and the upper bounds. In the ANN structure, there are input and output weights, and hence the output and input probability vectors ( $\mathbf{p}_h^O = [p_{h,1}^O, \dots, p_{h,M}^O]$  and  $\mathbf{p}_{hk}^I = [p_{hk,1}^I, \dots, p_{hk,M}^I]$ ) associated with output and input supports ( $\mathbf{z}_h^O = [z_{h,1}^O, \dots, z_{h,M}^O]$  and  $\mathbf{z}_{hk}^I = [z_{hk,1}^I, \dots, z_{hk,M}^I]$ ) are introduced in this reparameterization. Thus, I reparameterize  $\omega_h^O$  and  $\omega_{k,h}^I$  as

$$\begin{aligned} \omega_h^O &= \sum_{m=1}^M z_{h,m} p_{h,m}^O \\ \omega_{hk}^I &= \sum_{m=1}^M z_{hk,m} p_{hk,m}^I \end{aligned} \quad (4)$$

where  $p_{h,m}^O$  and  $p_{hk,m}^I$  are output and input probability estimates specified on the supports  $z_{h,m}$  and  $z_{hk,m}$  respectively. In terms of  $b_h^O$  and  $b_h^I$ , the reparameterization of these biases

is also somehow analogous to the weight parameter representation in probability and compact supports,

$$\begin{aligned} b_h^O &= \sum_{m=1}^M r_{h,m} q_{h,m}^O \\ b_h^I &= \sum_{m=1}^M r_{h,m} q_{h,m}^I \end{aligned} \quad (5)$$

where  $q_{h,m}^O$  and  $q_{h,m}^I$  are, respectively, the output and input probability estimates specified on the supports  $r_{h,m}$ . Just like the estimated weights and biases, the error  $\varepsilon_i$  is also viewed as the expected mean value of finite support  $v_i$ . Again, we can view error as the expected values of a random variable defined on a probability distribution. Thus,  $\varepsilon_i$  has a bounded support space  $\mathbf{v}_i = [\underline{v}_{i,1}, \dots, \bar{v}_{i,M}]$ , associated with  $i$  th observation, and weighted by the vector  $\mathbf{w}_i = [w_{i,1}, \dots, w_{i,M}]$ .

$$\varepsilon_i = \sum_{m=1}^M v_i w_{im}, \quad (6)$$

Pukelsheim (1994) suggested using the three-sigma rule for setting the support space of the error, such that  $\underline{v}_{i1} = -3\sigma$  and  $\bar{v}_{iM} = 3\sigma$ , where  $\sigma$  is the standard deviation of  $y$ . Now, the ANN model (Eq. 1) under the reparameterization becomes

$$y_i = \sum_{h=1}^H \left\{ \left( \sum_{m=1}^M z_{h,m} p_{h,m}^O \right) f^I \left( \sum_{k=1}^K \sum_{m=1}^M z_{hk,m} p_{hk,m}^I x_{i,k} + \sum_{m=1}^M r_{h,m} q_{h,m}^I \right) + \sum_{m=1}^M r_{h,m} q_{h,m}^O \right\} + \sum_{m=1}^M v_i w_{im}, \quad (7)$$

The entropy term is maximized subject to the requirements of the proper probability distributions for  $p_{h,m}^O$ ,  $p_{hk,m}^I$ ,  $q_{h,m}^O$ ,  $q_{h,m}^I$  and  $w_{i,m}$  and the  $N$  information-moment constraints of the ANN model. These unknown probabilities are assumed to be independent and can be estimated jointly by solving the constrained optimization problem with an objective function based on Shannon's entropy and constrains.

$$\begin{aligned} \mathbf{H}(\mathbf{p}^I, \mathbf{p}^O, \mathbf{q}^I, \mathbf{q}^O, \mathbf{w}) &= \underset{\mathbf{p}^I, \mathbf{p}^O, \mathbf{q}^I, \mathbf{q}^O, \mathbf{w}}{\operatorname{argmax}} \left\{ -\mathbf{H}(\mathbf{p}^I) - \mathbf{H}(\mathbf{p}^O) - \mathbf{H}(\mathbf{q}^I) - \mathbf{H}(\mathbf{q}^O) - \mathbf{H}(\mathbf{w}) \right\} \\ &= - \sum_{h=1}^H \sum_{k=1}^K \sum_{m=1}^M p_{hk,m}^I \log p_{hk,m}^I - \sum_{h=1}^H \sum_{k=1}^K \sum_{m=1}^M p_{hk,m}^O \log p_{hk,m}^O - \sum_{h=1}^H \sum_{m=1}^M q_{h,m}^I \log q_{h,m}^I \\ &\quad - \sum_{h=1}^H \sum_{m=1}^M q_{h,m}^O \log q_{h,m}^O - \sum_{i=1}^N \sum_{m=1}^M w_{im} \log w_{im} \end{aligned} \quad (8)$$

subject to Eq. (7) and additional contrarians

$$\sum_{m=1}^M z_{h,m} p_{h,m}^O = 1, \quad (9)$$

$$\sum_{m=1}^M z_{hk,m} p_{hk,m}^I = 1, \quad (10)$$

$$\sum_{m=1}^M r_{h,m} q_{h,m}^I = 1, \quad (11)$$

$$\sum_{m=1}^M r_{h,m} q_{h,m}^O = 1, \quad (12)$$

$$\sum_{m=1}^M v_i w_{im} = 1. \quad (13)$$

Then, the Lagrangian function is

$$\begin{aligned} \mathbf{L} = & -\mathbf{H}(\mathbf{p}^I) - \mathbf{H}(\mathbf{p}^O) - \mathbf{H}(\mathbf{q}^I) - \mathbf{H}(\mathbf{q}^O) - \mathbf{H}(\mathbf{w}) \\ & + \lambda^I \left[ y_i - \sum_{h=1}^H \left\{ \left( \sum_{m=1}^M z_{h,m} p_{h,m}^O \right) f^I \left( \sum_{k=1}^K \sum_{m=1}^M z_{hk,m} p_{hk,m}^I x_{i,k} + \sum_{m=1}^M r_{h,m} q_{h,m}^I \right) + \sum_{m=1}^M r_{h,m} q_{h,m}^O \right\} - \sum_{m=1}^M v_i w_{im} \right] \\ & + \rho \left[ 1 - \sum_{m=1}^M z_{h,m} p_{h,m}^O \right] + \Phi \left[ 1 - \sum_{m=1}^M z_{hk,m} p_{hk,m}^I \right] + \phi \left[ 1 - \sum_{m=1}^M r_{h,m} q_{h,m}^I \right] + \vartheta \left[ 1 - \sum_{m=1}^M r_{h,m} q_{h,m}^O \right] \\ & + \varphi \left[ 1 - \sum_{m=1}^M v_i w_{im} \right] \end{aligned} \quad (14)$$

The GME estimator generates the optimal probability vectors  $\widehat{\mathbf{p}}^I$ ,  $\widehat{\mathbf{p}}^O$ ,  $\widehat{\mathbf{q}}^I$ ,  $\widehat{\mathbf{q}}^O$  and  $\widehat{\mathbf{w}}$  that can be used to calculate point estimates of the unknown weights, biases and the unknown random errors through the reparameterizations in Eqs. (4-5), respectively. As noted by Golan et al. (1996), since the Lagrangian function (Eq. 14) is strictly concave, I can take the gradient of  $\mathbf{L}$  to derive the first-order conditions. I would like to note that the number of supports  $M$  is less controversial; and usually used in the literature is in the range between 3 and 7 points since there is likely no significant improvement in the estimation with more points in the support.

### 3 Experiment Study

In this section, I present the Monte Carlo simulations to illustrate the performance of ANN with GME estimation. More precisely, the suggested estimation is compared with the ML, Bayesian, and BP algorithms. In the case of GME, I set the number of support as 3 ( $M = 3$ ), whereas  $\mathbf{z}_k^O = \mathbf{z}_{hk}^I = \mathbf{r}_{h,m} = [-5, 0, 5]$  and  $\mathbf{v}_i = [-3(sd(y)), 0, 3(sd(y))]$ . For ML, the normal likelihood function is assumed, while the Gaussian approximation for the joint posterior probability distribution of the network weights and biases is assumed for Bayesian estimation. In the experiment, the output variable is simulated from

$$y_i = 1 + 0.5(\sin(0.5x_i)) + \varepsilon_i, \quad (15)$$

where  $\sin(\cdot)$  is the sinusoidal function. The simulated  $y_i$  becomes nonlinear and fluctuates overtime. Also, the precision of the estimations under different sample sizes and error distributions is to be investigated. Thus, I generated the error term from the normal and

non-normal distributions, consisting of  $N(0, 1)$ ,  $t(0, 1, 4)$ , and  $Unif(-1, 1)$ . Then, I generated a new sample during each Monte Carlo iteration by using Eq. (15) with the small sample sizes of 50 and 100 observations. The data are divided into training and test sets in which 70% of the total observations is used as the training data (in-sample data), while the rest is the test data (out-of-sample).

The simulation studies are carried out on a 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz, RAM 16 GB DDR5 workstation. The root mean square error (RMSE) is employed to report computation errors in all estimations. As there are several estimations considered and compared in this study, I set the same structure of ANN for all estimations. To be more specific, I set the learning rate  $\eta = 0.001$ , and the maximal error threshold 0.05. In addition, the single layer with sigmoid activation function is assumed, and the number of hidden neurons for those types of ANN models is set as 5. The above simulation process is repeated 100 times in order to estimate the mean value and standard deviation of RMSE (Table 1).

**Table 1.** Results of RMSE ( $n = 50$ )

In-sample	$\varepsilon_i \sim normal$			
	GME	BP	Bayesian	ML
Mean	0.814	0.668	0.670	0.660
SD	0.688	0.071	0.071	0.071
Out-of-Sample	$\varepsilon_i \sim normal$			
	GME	BP	Bayesian	ML
Mean	1.307	0.904	0.903	0.921
SD	1.123	0.166	0.170	0.165
In-sample	$\varepsilon_i \sim student - t$			
	GME	BP	Bayesian	ML
Mean	1.221	1.067	1.069	1.069
SD	1.273	0.115	0.116	0.115
Out-of-Sample	$\varepsilon_i \sim student - t$			
	GME	BP	Bayesian	ML
Mean	2.083	1.802	1.993	1.881
SD	2.114	0.693	0.893	0.701
In-sample	$\varepsilon_i \sim unif$			
	GME	BP	Bayesian	ML
Mean	2.117	2.652	2.745	2.784
SD	0.884	1.803	1.867	1.864
Out-of-Sample	$\varepsilon_i \sim unif$			
	GME	BP	Bayesian	ML
Mean	3.124	3.983	4.093	4.394
SD	1.093	2.343	2.431	2.993

Note: (1) The mean value of RMSE across 100 replications is reported, with the standard deviation in parentheses.

**Table 2.** Results of RMSE ( $n = 100$ )

In-sample	$\varepsilon_i \sim normal$			
	GME	BP	Bayesian	ML
Mean	0.743	0.535	0.573	0.544
SD	0.480	0.056	0.066	0.055
Out-of-Sample	$\varepsilon_i \sim normal$			
	GME	BP	Bayesian	ML
Mean	1.100	0.809	0.811	0.802
SD	1.023	0.123	0.136	0.126
In-sample	$\varepsilon_i \sim student - t$			
	GME	BP	Bayesian	ML
Mean	1.132	0.952	0.980	0.943
SD	1.341	0.327	0.207	0.321
Out-of-Sample	$\varepsilon_i \sim student - t$			
	GME	BP	Bayesian	ML
Mean	1.902	1.801	1.811	1.850
SD	2.493	0.955	0.907	0.939
In-sample	$\varepsilon_i \sim unif$			
	GME	BP	Bayesian	ML
Mean	2.334	5.685	5.693	5.383
SD	1.824	3.321	5.256	3.343
Out-of-Sample	$\varepsilon_i \sim unif$			
	GME	BP	Bayesian	ML
Mean	4.930	9.224	10.039	9.383
SD	2.549	4.003	5.023	5.034

Note: (1) The mean value of RMSE across 100 replications is reported, with the standard deviation in parentheses.

Reported in Tables 2, 3 are the mean and standard deviation of RMSE for in-sample goodness-of-fit and out-of-sample predictive accuracy across two horizons with three different error distributions. From these two tables, I can draw the following conclusions. (1) RMSE from the BP estimator is lower than that of the GME, ML, and Bayesian estimators, when the error of the ANN model is generated from normal and student-t distributions. The possible reason is that the small sample sizes of 50 and 100 may lead to a problem in the ML estimator as it relies on the asymptotic theory (Yamaka and Sriboonchitta 2020). Although the Bayesian estimation does not carry the assumptions of the asymptotic theory, which means that large sample size is not necessary for drawing valid statistical inferences, the conjugate prior for the weight parameter in this study may

not be well-specified and thereby leading to the higher RMSE than BP and ML. (2) When the error is assumed to be uniformly distributed, the GME estimator outperforms BP, Bayesian, and ML, because the mean of RMSE of the former is smaller than the latter. (3) With regard to the standard deviation, it is observed that the standard deviation from GME is relatively high in all error distributions, except uniform. This indicates that the variance of the GME is relatively high when the error distribution is known. However, it is also interesting to see that the proposed GME is superior to other estimations both in goodness-of-fit and predictive accuracy over all sample sizes, when the uniform error distribution is given. Therefore, in the case that the distribution of the error is unknown, the GME is considered a useful method as there is no need to assume the theoretical probability distribution for the errors to make statistical inference.

**Table 3.** Computational time (second) with different sample sizes

Method	Observations			
	50	100	500	1000
GME	19.139	35.993	104.335	904.024
BP	0.105	0.194	0.460	0.841
Bayesian	0.786	0.842	0.903	0.661
ML	0.203	0.225	0.509	0.798

Finally, it is interesting to assess the computational cost of each estimation for small and large sample sizes  $\{n = 50, 100, 500, 1000\}$ . It can be observed in Table 3 that GME spends 19.139s to 0.904.024s CPU time along 50 to 1000 observations. When comparing the computational performance between GME and other estimations, I found that GME runs slower than the others. This indicates that GME performs very poorly in the present simulations. This is not surprising due to the more parameters in the GME estimation. In other words, as the weight and bias parameters of ANN are derived from the expectation of probabilities on the prescribed supports, there will be a larger number of unknown parameters in the GME estimation. Although the GME takes high computational cost, it can provide more accurate prediction results particularly when the data is non-normally distributed.



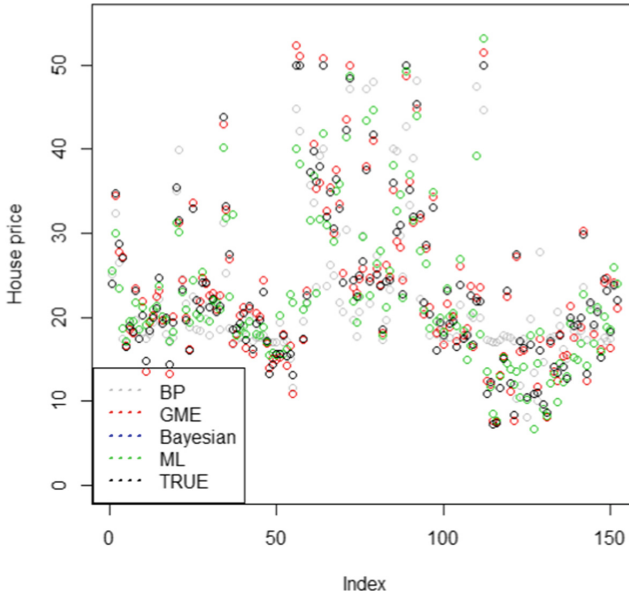
## 4 Case Study

Boston Housing is a dataset obtained from the UCI Machine Learning Repository. There are 506 observations for predicting the price of houses in Boston. The data contained 14 variables, consisting of 13 continuous variables (per capita crime rate by town, proportion of non-retail business acres per town, proportion of residential land zoned for lots over 25,000 sq.ft., nitrogen oxides pollutant concentration, average number of rooms, proportion of owner-occupied units built prior to 1940, weighted distances to five Boston employment centers, index of accessibility to radial highways, property-tax rate, pupil-teacher ratio by town, the proportion of blacks, percent lower status of the population and median house value) and one discontinuous variable (Charles river dummy variable). In this study, I consider median house value as output, while the rest are inputs. In the simulations, 354 training data and 152 testing data were randomly generated from the Boston Housing database.

**Table 4.** Forecast performance on the Boston housing data set

Estimation		RMSE
GME	In-sample	1.909
	Out-of-sample	2.839
BP	In-sample	2.632
	Out-of-sample	4.014
Bayesian	In-sample	4.623
	Out-of-sample	4.872
ML	In-sample	3.834
	Out-of-sample	3.993

The performance of each estimator is reported in Table 4. Note that the structure of ANN is assumed to be the same for all cases. With this study's focus on the improvement of the ANN estimation, the ANN having three layers and three hidden neurons is used. It can be seen that the GME has the lowest error out of the estimators compared in this real data study. The performance of the GME evaluated over the out-of-sample dataset is illustrated Fig. 1. It is clearly seen that the predicted values obtained from the GME estimator are close to the out-of-sample data. This indicates the high performance of the GME in estimating the ANN model.



**Fig. 1.** Fitting to test data

## 5 Conclusion

In this study, the GME estimator is suggested to be applied to ANN for its having several interesting and significant features different from the traditional estimators, namely BP, Bayesian, and ML. The estimator is effective in terms of goodness-of-fit and predictive ability by reparametrizing the weight and bias parameters as the expectation of random variables on the prescribed supports under the derived distributions of the entropy maximization, which is confirmed by the Monte Carlo simulations and real data example in this study. Moreover, using this estimator enables the production of a novel method capable of learning complex behaviors without human intervention and the model can be fitted without making specific assumptions. Therefore, I hypothesized that estimation with GME (GMS-ANN) would enable the resulting parameter estimates to be more unbiased to data distributions and robust to over-fitting issues compared to those of BP, ML, and Bayesian.

In order to compare the performance of GME and other competing estimators, the ANN structures are always assigned the same number of hidden neurons for both simulation and empirical studies. The RMSE is used for performance comparison. The results show that GME estimator produces the lowest RMSE estimates compared with BP, ML, and Bayesian when the errors are generated from uniform distributions. In other words, when the error distribution is unknown, these experiment results confirm an advantage of the GME approach. However, considering the computational cost, GME performs very poorly in the present simulations for all sample sizes due to the large number of probability estimates. It should be noted that in order to obtain as good performance as

possible for GME, long time effort is needed to find the appropriate probabilities for weight, bias, and error terms.

As the activation function in this study was assumed to be sigmoid, the performance of GME should be investigated considering other activation functions, such as exponential, ReLu and tanh. I leave this issue in the further study. Also, as the number of support and the value of bound can affect the estimation results, I would suggest varying the number and value of support bounds to validate the performance of GME in estimating ANN models.

## References

- Alibrandi, U., Mosalam, K.M.: Kernel density maximum entropy method with generalized moments for evaluating probability distributions, including tails, from a small sample of data. *Int. J. Numer. Meth. Eng.* **113**(13), 1904–1928 (2018)
- Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford (1995)
- Chon, K.H., Cohen, R.J.: Linear and nonlinear ARMA model parameter estimation using an artificial neural network. *IEEE Trans. Biomed. Eng.* **44**(3), 168–174 (1997)
- Chen, S., Mao, J., Chen, F., Hou, P., Li, Y.: Development of ANN model for depth prediction of vertical ground heat exchanger. *Int. J. Heat Mass Transf.* **117**, 617–626 (2018)
- Chen, B., Zhu, Y., Hu, J., Principe, J.C.: *System Parameter Identification: Information Criteria and Algorithms*. Newnes (2013)
- Chu, J., Liu, X., Zhang, Z., Zhang, Y., He, M.: A novel method overcome overfitting of artificial neural network for accurate prediction: application on thermophysical property of natural gas. *Case Stud. Therm. Eng.* **28**, 101406 (2021)
- Dorling, S.R., Foxall, R.J., Mandic, D.P., Cawley, G.C.: Maximum likelihood cost functions for neural network models of air quality data. *Atmos. Environ.* **37**(24), 3435–3443 (2003)
- Fu, L., Hsu, H.H., Principe, J.C.: Incremental backpropagation learning networks. *IEEE Trans. Neural Netw.* **7**(3), 757–761 (1996)
- Gish, H.: Maximum likelihood training of neural networks. In: *Artificial Intelligence Frontiers in Statistics*, pp. 241–255. Chapman and Hall/CRC (2020)
- Golan, A., Judge, G., Miller, D.: *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, Chichester (1996)
- Jaynes, E.T.: On the rationale of maximum-entropy methods. *Proc. IEEE* **70**(9), 939–952 (1982)
- Kocadağlı, O., Aşıkçıl, B.: Nonlinear time series forecasting with Bayesian neural networks. *Expert Syst. Appl.* **41**(15), 6596–6610 (2014)
- Kocadağlı, O.: A novel hybrid learning algorithm for full Bayesian approach of artificial neural networks. *Appl. Soft Comput.* **35**, 52–65 (2015)
- Lin, P., Fu, S.W., Wang, S.S., Lai, Y.H., Tsao, Y.: Maximum entropy learning with deep belief networks. *Entropy* **18**(7), 251 (2016)
- Maneejuk, P., Yamaka, W., Sriboonchitta, S.: Entropy inference in smooth transition kink regression. *Commun. Stat.-Simul. Comput.* 1–24 (2020)
- Müller, P., Insua, D.R.: Issues in Bayesian analysis of neural network models. *Neural Comput.* **10**(3), 749–770 (1998)
- Pukelsheim, F.: The three sigma rule. *Am. Stat.* **48**(2), 88–91 (1994)
- Ramos, V., Yamaka, W., Alorda, B., Sriboonchitta, S.: High-frequency forecasting from mobile devices' bigdata: an application to tourism destinations' crowdedness. *Int. J. Contemp. Hosp. Manag.* (2021)
- Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)

- Wang, X., Du, J., Wang, Y.: A maximum likelihood approach to deep neural network based speech dereverberation. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 155–158. IEEE (2017)
- White, H.: Some asymptotic results for learning in single hidden-layer feedforward network models. *J. Am. Stat. Assoc.* **84**(408), 1003–1013 (1989)
- Yamaka, W., Phadkantha, R., Maneejuk, P.: A convex combination approach for artificial neural network of interval data. *Appl. Sci.* **11**(9), 3997 (2021)
- Yamaka, W., Sriboonchitta, S.: Forecasting using information and entropy based on belief functions. *Complexity* **2020** (2020)
- Yang, Z., Baraldi, P., Zio, E.: A comparison between extreme learning machine and artificial neural network for remaining useful life prediction. In: 2016 Prognostics and System Health Management Conference (PHM-Chengdu), pp. 1–7. IEEE (2016)



# Robustness of Multi-criteria Nash Equilibrium Based on Vectorial Rationality Function

Urairat Deepan, Parin Chaipunya<sup>(✉)</sup>, and Poom Kumam

Department of Mathematics, Faculty of Science,  
King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd.,  
Bang Mod, Thung Khru, Bangkok 10140, Thailand  
urairat.deepan@mail.kmutt.ac.th, {parin.cha,poom.kum}@kmutt.ac.th

**Abstract.** This work concerns an economic model where numerous groups with certain objectives make decisions. We study a model  $M$  that is a parameterized class of “general games” together with an associated abstract vectorial rationality function. Finally, we prove that model  $M$  is structurally stable and robust to  $\epsilon$ -equilibria for “almost all” parameter values.

## 1 Formulation of the Problem

A theoretical framework for imagining social scenarios involving rival participants is game theory. In some ways, game theory can be seen as the science of strategy, or at the very least as the best possible way for independent, rival agents to make decisions in a strategic context.

The Nash equilibrium is a solution concept for a non-cooperative game, used here to describe the state where each agent cannot improve his decision without affecting the global cost. It is named after its inventor, John Forbes Nash Jr. [6], in 1950–1951. Suppose that  $I = \{1, \dots, N\}$  denotes the set of players,  $X = X_1 \times \dots \times X_i \times \dots \times X_N \subset \mathbb{R}^n$  is the action set of all players and  $X_i \subset \mathbb{R}$  is the action set of player  $i$ .

**Definition 1** [6]. Consider an  $N$ -player game where each player  $i$  minimizes the cost function  $F_i : X \rightarrow \mathbb{R}$ . A vector  $x^* = (x_i^*, x_{-i}^*) \in X$  is called a **Nash equilibrium** of this game if for every  $x_{-i}^*$  we have:

$$F_i(x_i^*, x_{-i}^*) \leq F_i(x_i, x_{-i}^*) \quad \forall x_i \in X_i, \quad \forall i \in I. \tag{1}$$

*Example 1.* Suppose that  $A$  and  $B$  are two players. Setting the cost function  $f_i : X_A \times X_B \rightarrow \mathbb{R}$  as below

For each  $i = \{A, B\}$ ,  $N = 2$ . Let  $x_A = \{\alpha, \beta\}$  and  $x_B = \{\alpha, \beta\}$ . From Table 1, we get that  $\bar{x} = (\beta, \beta)$  such that  $f_A(\beta, \beta) = +2 \leq f_A(\alpha, \beta) = +3$  and  $f_B(\beta, \beta) = +2 \leq f_B(\beta, \alpha) = +3$ .

**Table 1.** Cost function of players  $A$  and  $B$ .

	$B$	$\alpha$	$\beta$
$A$		$\alpha$	$\beta$
$\alpha$		+1	0
$\beta$		+1	+3
	$\alpha$	+3	+2
	$\beta$	0	+2

This work concerns an economic model where numerous groups with certain objectives make decisions. In order to produce the greatest and most accurate outcomes computationally, mathematicians observe that every group must choose a strategy that is unbiased and logical. However, this is difficult in practice as most of the groups would not agree with this perfect rationality condition. It is therefore interesting to seek for a robust model with a relaxed condition.

For this purpose, Anderlini and Canning [1] showed clearly that the assumption of perfect rationality is far too strict, they would like a model of bounded rationality as a basis for economic analysis. The authors established a model  $M$  that is a parametrized class of “general games” together with an associated abstract rationality function. Therefore, a model is structurally stable if the equilibrium set (given fully rational agents) varies continuously with the parameter values of the model.

Yu and Yu [8] proved that the set of critical parameter values is very small in the topological sense, and the model  $M$  is structurally stable and robust to  $\epsilon$ -equilibria for “almost all” parameter values.

In this paper, we propose the structural stability and robustness of multicriteria Nash equilibrium based on the vectorial rationality function. In Section 2, we formulate the mathematical model and also present key definitions and theories for the key outcomes. Finally the last section, we prove theorems for structurally stable and robustness of the model.

## 2 Mathematical Model

Let  $C \subseteq \mathbb{R}^n$  be a closed convex cone which is supposed pointed, that is,  $C \cap -C = \{0\}$  and to have a nonempty interior. The ordering  $\leq_C$  on  $\mathbb{R}^n$  is given by

$$y_1 \leq_C y_2 \Leftrightarrow y_2 - y_1 \in C. \tag{2}$$

For  $x, y \in \mathbb{R}^n$ , we shall write  $x <_C y$ , if  $y - x \in \text{Int } C$ .

Suppose that  $\mathbb{R}_+$  be a non-negative real valued. A model  $M$  consists of a quadruple  $(A, X, F, R)$ :  $A$  is the parameter space;  $X$  is the action space;  $F : A \times X \rightarrow 2^X$  is the feasibility set-valued mapping and  $F$  induces a further set-valued mapping  $f : A \rightarrow 2^X$ ,  $f(\lambda) = \{x \in X : x \in F(\lambda, x)\}$ ,  $\forall \lambda \in A$ ; the graph of  $f$ ,  $G(f) = \{(\lambda, x) \in A \times X : x \in f(\lambda)\}$ . Let positive real  $C = \mathbb{R}_+^n \subseteq \mathbb{R}^n$ .

$R : G(f) \rightarrow C$  is a rationality vector function,  $R(\lambda, x) = 0$  corresponds to the full rationality.

For all  $\lambda \in \Lambda, \epsilon \in \mathbb{R}_+^n$  the set of  $\epsilon$ -equilibria at  $\lambda$  is defined as

$$E(\lambda, \epsilon) = \{x \in f(\lambda) : R(\lambda, x) \in C \cap \overline{B_0}(\epsilon)\}, \tag{3}$$

where  $\overline{B_0}(\epsilon)$  represents the closed ball center at 0 with radius  $\epsilon$ .

In particular, the set of equilibria at  $\lambda$  is defined as

$$E(\lambda) = E(\lambda, 0) = \{x \in f(\lambda) : R(\lambda, x) = 0\}. \tag{4}$$

Throughout this paper,  $(\Lambda, \rho)$  is a complete metric space (may be noncompact),  $(X, d)$  is a compact metric space and  $E(\lambda) \neq \emptyset, \forall \lambda \in \Lambda$ .

Next, we will recall some definitions about the continuity of set-value mapping. Let  $X$  and  $Y$  be two metric spaces,  $K(X)$  be the set of all nonempty compact subsets of  $X$  and  $F : Y \rightarrow K(X)$  be a set-value mapping.

**Definition 2** [2]. Let  $X$  and  $Y$  be two metric spaces,  $K(X)$  be the set of all nonempty compact subsets of  $X$  and  $F : Y \rightarrow K(X)$  be a correspondence, then

1.  $F$  is said to be upper semicontinuous at  $y \in Y$  if for each open set  $U$  in  $X$  with  $U \supset F(y)$ , there exists an open neighborhood  $O(y)$  of  $y$  such that  $U \supset F(y')$  for each  $y' \in O(y)$ .
2.  $F$  is said to be lower semicontinuous at  $y$  if for each open set  $U$  in  $X$  with  $U \cap F(y) \neq \emptyset$ , there exists an open neighborhood  $O(y)$  of  $y$  such that  $U \cap F(y') \neq \emptyset$  for each  $y' \in O(y)$ .
3.  $F$  is said to be continuous at  $y$  if  $F$  is both upper and lower semicontinuous at  $y$ .

**Theorem 1** [2]. Let a set value mapping  $F : Y \rightarrow K(X)$ ,  $X$  be a compact metric space and  $Y$  be metric spaces,  $K(X)$  be the set of all nonempty compact subsets of  $X$ , then

1.  $F$  is said to be upper semi-continuous at  $y \in Y$  if and only if  $\forall y_n \rightarrow y, \forall x_n \in F(y_n), x_n \rightarrow x$ , then  $\exists x \in F(y)$ .
2.  $F$  is said to be lower semi-continuous at  $y$  if and only if  $\forall y_n \rightarrow y, \forall x \in F(y)$ , then  $\exists x_n \in F(y_n), x_n \rightarrow x$ .
3.  $F$  is said to be continuous at  $y$  if and only if for any  $y_n \rightarrow y, h(F(y_n), F(y)) \rightarrow 0$ , where  $h$  is the Hausdorff distance defined on  $Y$ .

**Definition 3** [8]. If for every  $\delta > 0, \exists \bar{\epsilon} > 0$  such that when  $\epsilon < \bar{\epsilon}$  and  $\rho(\lambda, \lambda') < \bar{\epsilon}, h(E(\lambda', \epsilon), E(\lambda')) < \delta$ , where  $h$  is the Hausdorff distance defined on  $X$ . Then the model  $M$  is robust to  $\epsilon$ -equilibria at  $\lambda \in \Lambda$ .

**Definition 4** [8]. The model  $M$  is structurally stable at  $\lambda \in \Lambda$  if the equilibrium set value mapping  $E : \Lambda \rightarrow K(X)$  is continuous at  $\lambda \in \Lambda$ .

Let  $\mathcal{A}$  is a metrizable space and  $C$  is a closed convex cone. We introduce the definition of vector function.

**Definition 5** [7]. A vector mapping  $T : \mathcal{A} \rightarrow \mathbb{R}^n$  is said to be  $C$ -lower semi-continuous at  $a \in \text{Dom } T = \{a \in \mathcal{A} | T(a) \in \mathbb{R}^n\}$ , if for any neighborhood  $V$  of zero in  $\mathbb{R}^n$ , there exists a neighborhood  $U$  of  $a$  such that

$$T(U) \subset T(a) + V + C. \tag{5}$$

**Definition 6** [3]. A vector mapping  $T : \mathcal{A} \rightarrow \mathbb{R}^n$  is said to be sequentially lower semicontinuous at  $a \in \text{Dom}T = \{a \in \mathcal{A} | T(a) \in \mathbb{R}^n\}$ , if for any  $b \in \mathbb{R}^n$  satisfying  $b \leq_C T(a)$  and for any sequence  $(\bar{a}_n)$  of  $\mathcal{A}$  which converges to  $a$ , there exists a sequence  $(b_n)$  in  $\mathbb{R}^n$  converging to  $b$  and satisfying  $b_n \leq_C T(\bar{a}_n)$ , for every  $n \in \mathbb{N}$

In [3], it has been proved that the two semicontinuity notions given in Definitions 5 and 6 do coincide whenever  $\mathcal{A}$  and  $\mathbb{R}^n$  are metrizable.

**Definition 7** [5]. A vector mapping  $T : \mathcal{A} \rightarrow \mathbb{R}^n$  is called quasi-lower semi-continuous or level closed if for every  $b \in \mathbb{R}^n$ ,

$$\text{lev}_{\leq_C b}(T) := \{\bar{a} \in \mathcal{A} | T(\bar{a}) \leq_C b\}$$

is closed.

*Remark 1* [5]. Notice that every epi-closed map is quasi-lower semi-continuous, while the converse is true if  $\text{Int } C \neq \emptyset$ . Also, every lower semi-continuous map is epi-closed but the converse may fail.

**Lemma 1** [4]. Let  $Y$  be a complete metric space,  $X$  be a metric space,  $F : Y \rightarrow K(X)$  be a upper semicontinuous correspondence. Then there exists a dense  $G_\delta$  subset  $Q$  of  $Y$  such that  $F$  is continuous at every  $y \in Q$ .

### 3 Analysis of the Problem and the Main Result

In this section, we assume the bounded rationality vector function and establish a theorem for structurally stable of models.

**Theorem 2.** Let  $f : \Lambda \rightarrow K(X)$  be an upper semi-continuous set-value mapping,  $R : G(f) \rightarrow C$  be a  $C$ -lower semi-continuous function. Then

- (1) the equilibrium correspondence  $E : \Lambda \rightarrow K(X)$  is upper semicontinuous;
- (2) there exists a dense countable intersection open sets  $G_\delta$  subset  $Q$  of  $\Lambda$  such that  $M$  is structurally stable at every  $\lambda \in Q$ .

*Proof 1.* (1) Since  $R$  is lower semi-continuous vector function, by Remark 1 we get that  $R$  is epi-closed, then  $R$  is quasi-lower semi-continuous or level closed which satisfies

$$\text{lev}_{\leq_C b}(R) := \{(\lambda, x) \in G(f) | R(\lambda, x) \leq_C b\}.$$

is closed by Definition 7. From  $E(\lambda) = \{x \in f(\lambda) | R(\lambda, x) \leq_C 0\}$ , so  $E(\lambda)$  is level closed. Since  $E(\lambda)$  is closed subset of compact set  $K(X)$  therefore  $E(\lambda)$  is compact.



We shall show that  $E : \Lambda \rightarrow K(X)$  is upper semicontinuous.  $\forall \lambda_n \rightarrow \lambda, \forall x_n \in E(\lambda_n), x_n \rightarrow x$ , then  $x_n \in f(\lambda_n)$  and  $R(\lambda_n, x_n) \leq_C 0$ . From  $f : \Lambda \rightarrow K(X)$  is upper semicontinuous, then  $x \in f(\lambda)$ .

Next, since  $R$  is  $C$ -lower semi-continuous at  $(\lambda, x)$ , for any  $b \in \mathbb{R}^n$  satisfying  $b \leq_C R(\lambda, x)$  and for any sequence  $(\lambda_n, x_n) \rightarrow (\lambda, x)$ , there exists a sequence  $b_n \rightarrow b$  satisfying  $b_n \leq_C R(\lambda_n, x_n), \forall n \in \mathbb{N}$ . Setting  $b = R(\lambda, x)$ , we get that

$$b_n \leq_C R(\lambda_n, x_n) \leq_C 0.$$

By ordering cone  $-b_n \in C$ . Since cone  $C$  is closed and  $b_n \rightarrow b$ , then  $-b = -R(\lambda, x) \in C$  this implies that  $R(\lambda, x) \in -C$ . Recall that  $R(\lambda, x) \in C$ . Since  $C$  is pointed,  $R(\lambda, x) \in C$  and  $R(\lambda, x) \in -C$ , then  $R(\lambda, x) \leq_C 0$ . Therefore  $x \in E(\lambda)$ .

(2) By Lemma 1, there exists a dense  $G_\delta$  subset  $Q$  of  $\Lambda$  such that the equilibrium correspondence  $E : \Lambda \rightarrow K(X)$  is continuous at every  $\lambda \in Q$ . Hence,  $M$  is structurally stable at every  $\lambda \in Q$ .

**Lemma 2 [9].** Let  $X$  and  $Y$  be two metric spaces,  $\{A_m\}_{m=1}^\infty$  be a sequence of  $K(X), \{y_m\}_{m=1}^\infty$  be a sequence of  $Y$  and  $\{f^m(x, y)\}_{m=1}^\infty$  be a sequence of continuous functions defined on  $X \times Y$ . If  $h(A_m, A) \rightarrow 0$ , where  $A \in K(X)$  and  $h$  is the Hausdorff distance defined on  $X, y_m \rightarrow y \in Y$  and

$$\sup_{(x,y) \in X \times Y} |f^m(x, y) - f(x, y)| \rightarrow 0$$

where  $f$  is a continuous function defined on  $X \times Y$ , then

$$\max_{w \in A_m} f^m(w, y_m) \rightarrow \max_{w \in A} f(w, y)$$

Under weaker assumptions than those in [1], we will show that structural stability implies robustness to  $\epsilon$ -equilibria.

**Theorem 3.** Under the assumptions of Theorem 2,  $M$  is structurally stable at  $\lambda \in \Lambda$  implies  $M$  is robust to  $\epsilon$ -equilibria at  $\lambda \in \Lambda$ .

*Proof 2.* Suppose that  $M$  were not robust to  $\epsilon$ -equilibria at  $\lambda \in \Lambda$ . Then there exists  $\delta > 0, \epsilon_n \rightarrow 0, \lambda_n \rightarrow \lambda$  such that

$$h(E(\lambda_n, \epsilon_n)) \geq \delta.$$

Since  $E(\lambda_n) \subset E(\lambda_n, \epsilon_n)$ , we can select  $x_n \in E(\lambda_n, \epsilon_n)$  such that

$$\min_{w \in E(\lambda_n)} d(x_n, w) > \frac{\delta}{2}.$$

Since  $X$  is compact, we may suppose without loss of generality that  $x_n \rightarrow x$ . Since  $E : \Lambda \rightarrow K(X)$  is continuous at  $\lambda$ , then  $h(E(\lambda_n), E(\lambda)) \rightarrow 0$ . By Lemma 2, we get

$$\min_{w \in E(\lambda)} d(x, w) \geq \frac{\delta}{2}.$$

Since  $x_n \in f(\lambda_n)$  and  $f : \Lambda \rightarrow K(X)$  is upper semicontinuous, then  $x \in f(\lambda)$ . Since  $R(\lambda_n, x_n) \in C \cap \overline{B_0}(\epsilon_n)$  and vector function  $R$  is lower semicontinuous at  $(\lambda, x)$  by Definition 6, then  $R(\lambda, x) = 0$ ,  $x \in E(\lambda)$  which contradicts the fact that  $\min_{w \in E(\lambda)} d(x, w) \geq \frac{\delta}{2}$ . Hence,  $M$  must be robust to  $\epsilon$ -equilibria at  $\lambda \in \Lambda$ .

**Acknowledgments.** The authors acknowledge the financial support provided by the Center of Excellence in Theoretical and Computational Science (TaCS-CoE), KMUTT. The first author appreciates the support provided by Petchra Pra Jom Klao Ph.D. Research Scholarship through grant no (17/2564), by King Mongkut's University of Technology Thonburi.

## References

1. Anderlini, L., Canning, D.: Structural stability implies robustness to bounded rationality. *J. Econ. Theory* **101**(2), 395–422 (2001)
2. Aubin, J.-P., Frankowska, H.: *Set-Valued Analysis*. Springer, Heidelberg (2009). <https://doi.org/10.1007/978-0-8176-4848-0>
3. Combari, C., Laghdir, M., Thibault, L.: Sous-différentiels de fonctions convexes composées. *Ann. Sci. Math. Québec* **18**(2), 119–148 (1994)
4. Fort, M.K.: Points of continuity of semicontinuous functions. *Publ. Math. Debrecen* **2**(1951), 100–102 (1951)
5. Mansour, M., Metrane, A., Thera, M.: Lower semicontinuous regularization for vector-values mappings: rapport de recherche (2004)
6. Nash Jr., J.F.: Equilibrium points in  $n$ -person games. *Proc. Natl. Acad. Sci.* **36**(1), 48–49 (1950)
7. Théra, M.: Etude des fonctions convexes vectorielles semi-continues. Doctorat de Troisième Cycle (76) (1978)
8. Yu, C., Yu, J.: On structural stability and robustness to bounded rationality. *Non-linear Anal.: Theory Methods Appl.* **65**(3), 583–592 (2006)
9. Yu, J.: Essential equilibria of  $n$ -person noncooperative games. *J. Math. Econ.* **31**(3), 361–372 (1999)



# Why Quantiles Are a Good Description of Volatility in Economics: An Alternative Explanation

Laxman Bokati<sup>1</sup>, Olga Kosheleva<sup>2</sup>, Vladik Kreinovich<sup>3</sup>(✉),  
and Kittawit Autchariyapanitkul<sup>4</sup>

<sup>1</sup> Computational Science Program, University of Texas at El Paso,  
500 W. University, El Paso, TX 79968, USA

lbokati@miners.utep.edu

<sup>2</sup> Department of Teacher Education, University of Texas at El Paso,  
500 W. University, El Paso, TX 79968, USA

olgak@utep.edu

<sup>3</sup> Department of Computer Science, University of Texas at El Paso,  
500 W. University, El Paso, TX 79968, USA

vladik@utep.edu

<sup>4</sup> Faculty of Economics,

Maejo University, Chiang Mai, Thailand

kittawit.a@mju.ac.th

**Abstract.** In econometrics, volatility of an investment is usually described by its Value-at-Risk (VaR), i.e., by an appropriate quantile of the corresponding probability distribution. The motivations for selecting VaR are largely empirical: VaR provides a more adequate description of what people intuitively perceive as risk. In this paper, we analyze this situation from the viewpoint of decision theory, and we show that this analysis naturally leads to the Value-at-Risk, i.e., to a quantile.

Interestingly, this analysis also naturally leads to an optimization problem related to quantile regression.

## 1 Description of the Problem

**Need to Represent a Random Gain by a Single Number.** In economics, the outcomes of a decision are usually known with uncertainty. Based on the previous experience, for each possible decision, we can estimate the probability of different gains  $m$ . Thus, each possible decision can be characterised by a probability distribution on the set of possible gains  $m$ . This probability distribution can be described by a probability density function  $\rho(m)$ , or by the cumulative distribution function  $F(n) \stackrel{\text{def}}{=} \text{Prob}(m \leq n)$ .

To select the best decision, we need to be able to compare every two possible decisions – and for this purpose, we need to represent each possible decision by a single number.

**Problem: What Should this Number be?** How can we select this number?

According to decision theory, decisions of a rational person are equivalent to selecting a decision that leads to the largest possible mean value of this person's utility (see,

e.g., [3,4,7–11]), and in the first approximation, utility is proportional to the gain. According to this logic, we should select a decision that leads to the largest possible value of the mean gain.

**What We Do in this Paper.** In this paper, we show that a more appropriate decision is to select the decision with the largest possible value of the appropriate quantile. This provides an additional explanation for the fact that in econometrics, an appropriate quantile – known as the Value at Risk (VaR) (see, e.g., [2]) – is an accepted measure of the investment’s volatility (for other explanations, see, e.g., [1]).

## 2 Analysis of the Problem and Its Resulting Formulation in Precise Terms

Suppose that we represent a decision by a number  $n$ . Since the outcomes are random, the actual gain  $m$  will be, in general, different from  $n$ . How will this difference affect the decision maker?

**Case When We Gained More than Expected.** Let us first consider the case when the actual gain  $m$  is larger than  $n$ . In this case, we can use the unexpected surplus  $m - n$ . For example, a person can take a trip, a company can buy some new equipment, etc. However, the value of this additional amount to the user is somewhat decreased by the fact that this amount was unexpected. For example, if a user plans a trip way beforehand, it is much cheaper than buying it in the last minute. If the company plans to buy an equipment some time ahead, it can negotiate a better price. In all these cases, in comparison to the user’s value of each dollar of the expected amount  $n$ , each dollar from the unexpected additional amount  $m - n$  has a somewhat lower value, valued less by some coefficient  $\alpha_+ > 0$ .

The loss of value for each dollar above the expected value is  $\alpha_+$ . Thus, the overall loss corresponding to the whole unexpected amount  $m - n$  is equal to  $\alpha_+ \cdot (m - n)$ . So, to get the overall user’s value  $v(m)$  of the gain  $m$ , we need to subtract this loss from  $m$ :

$$v(m) = m - \alpha_+ \cdot (m - n). \quad (1)$$

**Case When We Gained Less than Expected.** What if the actual gain  $m$  is smaller than  $n$ ? In this case, not only we lose the difference  $n - m$  in comparison to what we expected, but we lose some more. For example, since we expected the gain  $n$ , we may have already made some purchases for which we planned to pay from this amount. Since we did not get as much money as we expected, we need to borrow the missing amount of money – and since borrowing money comes with an interest, we thus lose – e.g., on this interest – some additional amount. In other words, for each dollar that we did not receive, we lose some additional amount; let us denote this additional loss by  $\alpha_- > 0$ .

The overall additional loss to the user caused by the difference  $n - m$  can be obtained by multiplying this difference by the per-dollar loss  $\alpha_-$ , so this loss is equal to

$\alpha_- \cdot (n - m)$ . The overall user's value  $v(m)$  corresponding to the gain  $m$  can be thus obtained by subtracting this loss from the monetary amount  $m$ :

$$v(m) = m - \alpha_- \cdot (n - m). \quad (2)$$

**Resulting Optimization Problem.** As we have mentioned, a rational agent should maximize the expected value, i.e., a rational agent should select the value  $n$  that maximizes the expression

$$\begin{aligned} V &\stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \rho(m) \cdot v(m) dm \\ &= \int_{-\infty}^n \rho(m) \cdot (m - \alpha_- \cdot (n - m)) dm + \int_n^{\infty} \rho(m) \cdot (m - \alpha_+ \cdot (m - n)) dm. \end{aligned} \quad (3)$$

### 3 Solving the Resulting Optimization Problem

**Solving the Problem.** Differentiating the expression (3) with respect to the unknown value  $n$  and equating the resulting derivative to 0, we conclude that

$$-\alpha_- \cdot \int_{-\infty}^n \rho(m) dm + \alpha_+ \cdot \int_n^{\infty} \rho(m) dm = 0 \quad (4)$$

Here, by definition of the probability density:

$$\int_{-\infty}^n \rho(m) dm = \text{Prob}(m \leq n) = F(n)$$

and

$$\int_n^{\infty} \rho(m) dm = \text{Prob}(m \geq n) = 1 - F(n).$$

Thus, the equality (4) takes the form

$$-\alpha_- \cdot F(n) + \alpha_+ \cdot (1 - F(n)) = 0,$$

so

$$(\alpha_- + \alpha_+) \cdot F(n) = \alpha_+$$

and hence

$$F(n) = \frac{\alpha_+}{\alpha_- + \alpha_+}.$$

This is exactly the quantile corresponding to

$$\tau = \frac{\alpha_+}{\alpha_- + \alpha_+}. \quad (5)$$

Thus we arrive at the following conclusion.

**Conclusion.** In the above natural optimization problem, the optimal value  $n$  representing the random variable described by a cumulative distribution function  $F(m)$  is the quantile corresponding to the value (5).

This explains why quantiles (i.e., VaR) indeed work well in econometrics.

**An Interesting Observation.** In econometrics, quantiles are not only used to describe the risk of different investments, they are also used to describe the dependence between different random variables – in the form of describing how the quantile of the dependent variable  $m$  depends on the quantiles of the corresponding independent variables  $x_1, \dots, x_n$ . In this technique – known as *quantile regression* (see, e.g., [5, 6]), for each value  $\tau \in (0, 1)$  the  $\tau$ -level quantile  $n$  of the random variable  $m$  is determined by minimizing the expression

$$I \stackrel{\text{def}}{=} (\tau - 1) \cdot \int_{-\infty}^n \rho(m) \cdot (m - n) dm + \tau \cdot \int_n^{\infty} \rho(m) \cdot (m - n) dm. \quad (6)$$

By comparing the formulas (3) and (7) for the value  $\tau$  determined by the formula (5), we see that

$$V = \int_{-\infty}^{\infty} \rho(m) \cdot m dm - (\alpha_+ + \alpha_-) \cdot I. \quad (7)$$

The first integral in the expression (7) is just the expected value of the gain, it does not depend on  $n$  at all. Thus, maximizing  $V$  is equivalent to minimizing the expression  $I$ .

So, the formal optimized expression used in quantile regression actually has a precise meaning: it is linearly related to the expected utility of the user.

**Acknowledgments.** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

This work was also supported by the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand.

## References

1. Aguilar, S., Kreinovich, V., Pham, U.: Why quantiles are a good description of volatility in economics: a pedagogical explanation. In: Sriboonchitta, S., Kreinovich, V., Yamaka, W. (eds.) TES 2022. Studies in Systems, Decision and Control, vol. 429, pp. 3–6. Springer, Cham (2023). [https://doi.org/10.1007/978-3-030-97273-8\\_1](https://doi.org/10.1007/978-3-030-97273-8_1)
2. Auer, M.: Hands-On Value-at-Risk and Expected Shortfall: A Practical Primer. Springer, Heidelberg (2018). <https://doi.org/10.1007/978-3-319-72320-4>
3. Fishburn, P.C.: Utility Theory for Decision Making. Wiley, New York (1969)
4. Fishburn, P.C.: Nonlinear Preference and Utility Theory. The John Hopkins Press, Baltimore (1988)
5. Furno, M., Vistocco, D.: Quantile Regression: Estimation and Simulation. Wiley, Hoboken (2018)
6. Koenker, R., Chernozhukov, V., He, X., Peng, L. (eds.): Handbook of Quantile Regression. Chapman & Hall/CRC, Boca Raton (2017)
7. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) Human-Centric Decision-Making Models for Social Sciences. SCI, vol. 502, pp. 163–193. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-642-39307-5\\_8](https://doi.org/10.1007/978-3-642-39307-5_8)

8. Luce, R.D., Raiffa, R.: *Games and Decisions: Introduction and Critical Survey*. Dover, New York (1989)
9. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: Decision making beyond arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction. *Int. J. Intell. Syst.* **24**(1), 27–47 (2009)
10. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics Under Interval and Fuzzy Uncertainty*. Springer, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-24905-1>
11. Raiffa, H.: *Decision Analysis*. McGraw-Hill, Columbus (1997)



# Hawthorne Effect: An Explanation Based on Decision Theory

Sofia Holguin<sup>1</sup>, Vladik Kreinovich<sup>1(✉)</sup>, and Phuong Hoang Nguyen<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Texas at El Paso,  
500 W. University, El Paso, TX 79968, USA  
seholguin2@miners.utep.edu, vladik@utep.edu

<sup>2</sup> Artificial Intelligence Division, Information Technology Faculty, Thang Long University,  
Nghiem Xuan Yem Road, Hoang Mai District, Hanoi, Vietnam

**Abstract.** It is known that people feel better (and even work better) if someone pays attention to them; this is known as the *Hawthorne effect*. At first glance, it sounds counter-intuitive: this attention does not bring you any material benefits, so why would you feel better? If you are sick and someone gives you medicine, this will make you feel better, but if someone just pays attention, why does that make you feel better? In this paper, we use the general ideas of decision theory to explain this seemingly counterintuitive phenomenon.

## 1 Formulation of the Problem

**What is Hawthorne Effect.** If someone helps a person, this usually makes this person happier. Interestingly, if this someone does not actually help, but simply expresses some interest in this person's problem, this also makes the person happier. For example, when a research team comes to study not-very-comfortable working conditions, this very attention already makes the workers happier – and even boosts their productivity, although the working conditions have not improved and there is no specific plan to improve them. Similarly, the very attention to a sick person makes this person feel better, even though this attention does not lead to any improvement of the health situation. This phenomenon was first documented in a factory called Hawthorne Works. Because of this fact, this effect is known as the *Hawthorne effect*; see, e.g., [1] and references therein.

It should be mentioned that the feeling-better phenomenon only occurs when the people paying attention have a positive attitude toward the folks they pay attention to. Definitely, if a team would come to analyze the workers for the potential purpose of making them work harder for the same pay, this attention would not make the workers feel better or be more productive.

**Why Hawthorne Effect?** At first glance, this phenomenon sounds counter-intuitive: why would workers feel better if some strangers whom they see for the first time and probably not see again simply study their working conditions? Ok, people crave for attention, but the increase in happiness and productivity is disproportionate: it is comparable to a similar increase caused by the actual improvement in the working conditions.



How can we explain this seemingly counter-intuitive phenomenon? In this paper, we analyze this situation from the viewpoint of decision theory and show that, within this theory, the Hawthorne effect can indeed be naturally explained.

## 2 Our Explanation

**Decision Theory: A Brief Reminder.** To come up with the desired explanation, let us recall the main ideas behind decision theory; see, e.g., [2–5, 8–10]. Decision theory studies decisions made by rational people, i.e., people whose preferences are consistent: e.g., if a person prefers an alternative  $A$  to alternative  $B$  and prefers alternative  $B$  to some other alternative  $C$ , then, when presented with a choice between  $A$  and  $C$ , this person should select  $A$ .

It turns out that under such consistency conditions, decisions of such a rational person can be described by a number-valued function  $u(A)$  called *utility* so that out of several alternatives  $A_1, \dots, A_n$ , the person always selects an alternative with the largest possible value of utility.

A person's utility may depend not only on the objective circumstances of this person, but also on the utilities of others. This dependence is usually described by a linear formula:

$$u_i = u_i^{(0)} + \sum_j \alpha_{ij} \cdot u_j,$$

where  $u_i^{(0)}$  is the utility corresponding to the person  $i$ 's objective circumstances, and the coefficients  $\alpha_{ij}$  describe  $i$ 's attitude towards person  $j$ ; see, e.g., [8] and references therein.

For collective decision making, the optimal solution – according to decision theory – is to maximize the product of utilities; this is known as *Nash's bargaining solution*; see, e.g., [5–7].

**Let us Apply Decision Theory to Our Situation.** Let us consider a simplified situation, in which we have two persons: the main Person 1 and another Person 2 who starts expressing interest in Person 1.

We want to describe how this interest affects the happiness of Person 1. In general, according to decision theory, this happiness is determined by the overall decision  $a$ . In general, the state of each system – and, in particular, each decision – can be described by providing numerical values of all the characteristics describing this state or this decision. So, in mathematical terms, each decision can be described by a tuple of the corresponding numerical values  $a = (a_1, \dots, a_k, \dots)$ .

At first, before the interest starts, the collective decision  $f$  is determined by maximizing the product of the utilities of these folks:

$$u_1(f) \cdot u_2(f) = \max_a u_1(a) \cdot u_2(a). \quad (1)$$

Once the Person 2 starts getting positively interested in Person 1, the utility of Person 2 changes from its original value  $u_2(a)$  to the new value  $u_2(a) + \alpha \cdot u_1(a)$  for some positive number  $\alpha > 0$ . We consider the case when this interest is mostly professional,

so its intensity is not high:  $\alpha \ll 1$ . Since the utility of Person 2 changes, the collective solution also changes, now we select an alternative  $s$  that maximizes the product of new utilities:

$$u_1(s) \cdot (u_2(s) + \alpha \cdot u_1(s)) = \max_a u_1(a) \cdot (u_2(a) + \alpha \cdot u_1(a)). \tag{2}$$

In these terms, the Hawthorne effect means that this interest makes Person 1 happier, i.e., that

$$u_1(s) > u_1(f). \tag{3}$$

Let us see if we can explain this effect.

**Towards an Explanation.** According to calculus, the fact that the expression (1) attains its maximum for  $a = f$  means that the derivatives of this expression over all components  $a_k$  of  $a$  are equal to 0:

$$\frac{\partial(u_1(a) \cdot u_2(a))}{\partial a_k} \Big|_{a=f} = 0. \tag{4}$$

At the point  $a = f$ , the derivative  $d_k$  of the new objective function

$$u_1(a) \cdot (u_2(a) + \alpha \cdot u_1(a)) = u_1(a) \cdot u_2(a) + \alpha \cdot (u_1(a))^2 \tag{5}$$

with respect to the component  $a_k$  is equal to

$$\begin{aligned} d_k &\stackrel{\text{def}}{=} \frac{\partial(u_1(a) \cdot (u_2(a) + \alpha \cdot u_1(a)))}{\partial a_k} \Big|_{a=f} \\ &= \frac{\partial(u_1(a) \cdot u_2(a) + \alpha \cdot (u_1(a))^2)}{\partial a_k} \Big|_{a=f} \\ &= \frac{\partial(u_1(a) \cdot u_2(a))}{\partial a_k} \Big|_{a=f} + \frac{\partial(\alpha \cdot (u_1(a))^2)}{\partial a_k} \Big|_{a=f}. \end{aligned} \tag{6}$$

At the point  $a = f$ , the first term in the sum (6) is, according to the formula (4), equal to 0, so

$$d_k = \frac{\partial}{\partial a_k} (\alpha \cdot (u_1(a))^2) \Big|_{a=f} = 2\alpha \cdot u_1(f) \cdot \frac{\partial u_1(a)}{\partial a_k} \Big|_{a=f}. \tag{7}$$

Depending of the sign of the last derivative in the formula (7), we have two possible cases: either this derivative is positive (or, strictly speaking, non-negative) or it is negative. Let us consider both cases one by one.

**Case When the Derivative is Positive.** Let us first consider the case when the derivative is positive, i.e., when

$$\frac{\partial u_1(a)}{\partial a_k} \Big|_{a=f} > 0. \tag{8}$$

Since  $u_1(f) > 0$  and  $\alpha > 0$ , this means that the derivative  $d_k$  is also positive – so, if we increase the value  $a_k$ , we get a larger value of the product of the utilities. So, to get to the new maximum  $s$ , we need to increase  $a_k$ .

In this case, due to (8), the value of the utility  $u_1(a)$  will also increase – i.e., in commonsense terms, Person 1 will be happier in the new state  $s$  than in the original state  $s$ .

**Case When the Derivative is Negative.** Let us now consider the case when the derivative is negative, i.e., when

$$\frac{\partial u_1(a)}{\partial a_k} \Big|_{a=f} < 0. \quad (9)$$

Since  $u_1(f) > 0$  and  $\alpha > 0$ , this means that the derivative  $d_k$  is also negative – so, if we decrease the value  $a_k$ , we get a larger value of the product of the utilities. So, to get to the new maximum  $s$ , we need to decrease  $a_k$ .

In this case, due to (9), the value of the utility  $u_1(a)$  will also increase – i.e., in commonsense terms, Person 1 will be happier in the new state  $s$  than in the original state  $s$ .

**Conclusion.** In both cases, the mere fact that Person 2 starts expressing interest in Person 1 increases the happiness level of Person 1 – which is exactly the Hawthorne effect.

Thus, this seemingly counter-intuitive effect indeed naturally follows from decision theory.

**Acknowledgments.** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Draper, S.W.: The Hawthorne, Pygmalion, placebo, and other expectancy effects: some notes. <http://www.psy.gla.ac.uk/~steve/hawth.html>. Accessed May 2022
2. Fishburn, P.C.: *Utility Theory for Decision Making*. Wiley, New York (1969)
3. Fishburn, P.C.: *Nonlinear Preference and Utility Theory*. The John Hopkins Press, Baltimore (1988)
4. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) *Human-Centric Decision-Making Models for Social Sciences*. SCI, vol. 502, pp. 163–193. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-642-39307-5\\_8](https://doi.org/10.1007/978-3-642-39307-5_8)
5. Luce, R.D., Raiffa, R.: *Games and Decisions: Introduction and Critical Survey*. Dover, New York (1989)
6. Nash, J.: The bargaining problem. *Econometrica* **18**(2), 155–162 (1950)
7. Nguyen, H.P., Bokati, L., Kreinovich, V.: New (simplified) derivation of Nash’s bargaining solution. *J. Adv. Comput. Intell. Inform. (JACIII)* **24**(5), 589–592 (2020)
8. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: Decision making beyond arrow’s ‘impossibility theorem’, with the analysis of effects of collusion and mutual attraction. *Int. J. Intell. Syst.* **24**(1), 27–47 (2009)
9. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics Under Interval and Fuzzy Uncertainty*. Springer, Berlin, Heidelberg (2012). <https://doi.org/10.1007/978-3-642-24905-1>
10. Raiffa, H.: *Decision Analysis*. McGraw-Hill, Columbus (1997)



# Fair Bankruptcy Solutions Under Interval Uncertainty

Uyen Pham<sup>1</sup>, Olga Kosheleva<sup>2</sup>, and Vladik Kreinovich<sup>2</sup>(✉)

<sup>1</sup> University of Economics and Law, Ho Chi Minh City, Vietnam  
uyenph@uel.edu.vn

<sup>2</sup> University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA  
{olgak, vladik}@utep.edu

**Abstract.** If the overall amount of the company's assets is smaller than its total debts, then a fair solution is to give, to each creditor, the amount proportional to the corresponding debt, e.g., 10 cents for each dollar or 50 cents for each dollar. But what if the debt amounts are not known exactly, and for some creditors, we only know the lower and upper bounds on the actual debt amount? What division will be fair in such a situation? In this paper, we show that the only fair solution is to make payments proportional to an appropriate convex combination of the bounds – which corresponds to Hurwicz optimism-pessimism criterion for decision making under interval uncertainty.

## 1 Formulation of the Problem

**What is a Bankruptcy Problem.** A company goes bankrupt if the total amount of its assets is smaller than the total amount of debts. Some of the debts have priority – e.g., according to the US labor law, salary needs to be paid in full, irrespective of debts to others. Once these priority debts are paid, we face a problem of how to divide the remaining assets  $A$  between the creditors to whom the company owes amounts  $d_1, \dots, d_n$ .

**How this Problem is Usually Solved.** In this case, a usual solution is to make payments proportional to debts, i.e., depending on the ratio between the assets and the debts, 10 cents per dollar, 50 cents per dollar, etc. In general, the amount  $g_i$  given to the  $i$ -th creditor is equal to

$$g_i = d_i \cdot \frac{A}{\sum_{j=1}^n d_j}. \quad (1)$$

**Need to Take Interval Uncertainty into Account.** In some cases, the debt is purely monetary, and its amount  $d_i$  is known exactly. In many practical situations, however, the situation is more complicated, so for many creditors, we only know the bounds  $\underline{d}_i \leq d_i \leq \bar{d}_i$  of the actual debt amount. How should we divide the assets in this situation?

**Case of Interval Uncertainty: How is this Problem Solved Now.** Several papers describe how to solve the bankruptcy problem under interval uncertainty. For example, the paper [2] suggests selecting a single value  $d_i$  within each interval, and then

using these values  $d_i$  to divide the assets. For example, to select  $d_i$ , we can use Hurwicz optimism-pessimism criterion [3, 6, 8]: namely, we agree on some value  $\alpha \in [0, 1]$  and take  $d_i = \alpha \cdot \bar{d}_i + (1 - \alpha) \cdot \underline{d}_i$ .

A more complex scheme was proposed in [7] – following a solution to a similar problem in [15].

**What We do in this Paper.** In this paper, we show that a natural formalization of fairness uniquely determines Hurwicz-based solutions – which are thus recommended as the fair ones.

## 2 How to Describe Fairness

**Fairness: First Requirement.** Fairness means, first, that if the debt  $d_i$  to creditor  $i$  is smaller than or equal to the debt  $d_j$  to creditor  $j$ , then the payment  $g_i$  to creditor  $i$  should be smaller than or equal to the payment to creditor  $j$ .

**Fairness: Second Requirement.** Second, fairness means that two creditors should not gain or lose by joining together. In other words:

- if for debts  $d_1, d_2, d_2, \dots, d_n$ , we had payments  $g_1, g_2, g_3, \dots, g_n$ ,
- them for debts  $d_1 + d_2, d_2, \dots, d_n$ , we should have payments  $g_1 + g_2, g_3, \dots, g_n$ .

**Continuity.** It also makes sense to require that if in two situations, debts are close, then payments should be close – i.e., that payments should be a continuous function of debts.

## 3 What if We Impose Fairness Requirements in Situations When We Know the Exact Amount of Debts

Before we consider the case of interval uncertainty, let us analyze what will happen if we impose fairness requirements in the situations when we know the exact amount of debt.

**Definition 1.** Let  $A < D$  be two positive numbers.

- We will call  $A$  the amount of assets, and we will call  $D$  the amount of debt.
- By a solution to the bankruptcy problem (or simply solution, for short), we mean a function  $S$  that maps every tuple  $\langle d_1, \dots, d_n \rangle$  of positive real numbers for which  $d_1 + \dots + d_n = D$  into a tuple of non-negative real numbers  $\langle g_1, \dots, g_n \rangle$  for which

$$g_1 + \dots + g_n = A.$$

**Definition 2.** We say that the solution  $S$  is fair if it satisfies the following two requirements for each tuple  $\langle d_1, d_2, d_3, \dots, d_n \rangle$  and for  $S(\langle d_1, \dots, d_n \rangle) = \langle g_1, \dots, g_n \rangle$ :

- if  $d_i \leq d_j$ , then  $g_i \leq g_j$ ;
- $S(\langle d_1 + d_2, d_3, \dots, d_n \rangle) = \langle g_1 + g_2, g_3, \dots, g_n \rangle$ .

**Definition 3.** We say that the solution  $S$  is continuous, if for every  $n$ , if  $d_i^{(k)} \rightarrow d_i$  for all  $i$ ,  $S(\langle d_1^{(k)}, \dots, d_n^{(k)} \rangle) = \langle g_1^{(k)}, \dots, g_n^{(k)} \rangle$ , and  $g_i^{(k)} \rightarrow g_i$  for all  $i$ , then

$$S(\langle d_1, \dots, d_n, A \rangle) = \langle g_1, \dots, g_n \rangle.$$

**Proposition 1.** For each solution  $S$ , the following two conditions are equivalent to each other:

- the solution is fair and continuous,
- the solution has the form

$$g_i = d_i \cdot (A/D). \quad (2)$$

*Comment.* So, the usual solution is the only one which is fair (and continuous).

**Proof.** It is easy to check that the above solution is fair and continuous. So, to complete the proof, it is sufficient to prove that every fair continuous solution  $S$  has this form.

Indeed, let  $S$  be a fair and continuous solution. For every natural number  $N$ , we can consider the tuple  $\langle d_1, \dots, d_N \rangle = \langle D/N, \dots, D/N \rangle$  consisting of  $N$  equal debt values. By the first fairness requirement, since the debts  $d_i$  are all equal, the payments  $g_i$  are also all equal. Since  $g_1 + \dots + g_N = A$ , this means that  $N \cdot g_i = A$  hence  $g_i = A/N$ , and the payments tuple has the form  $\langle g_1, \dots, g_N \rangle = \langle A/N, \dots, A/N \rangle$ .

For any sequence of natural numbers  $k_1, \dots, k_n$  for which  $k_1 + \dots + k_n = N$ , the tuple  $\langle k_1 \cdot (D/N), \dots, k_n \cdot (D/N) \rangle$  can be obtained from the tuple  $\langle 1/N, \dots, 1/N \rangle$  by adding up the first  $k_1$  terms, then the next  $k_2$  terms, etc. So, due to the second fairness requirements, the resulting payment tuple  $\langle g_1, \dots, g_n \rangle$  can be obtained from the tuple  $\langle A/N, \dots, A/N \rangle$  by adding the first  $k_1$  terms, then the next  $k_2$  terms, etc. Thus, the payment tuple has the form  $\langle k_1 \cdot (A/N), \dots, k_n \cdot (A/N) \rangle$ . In other words, for each debt  $d_i = k_i \cdot (D/N)$ , the payment is equal to  $g_i = k_i \cdot (A/N)$ . From  $d_i = k_i \cdot (D/N)$ , we conclude that  $k_i = d_i \cdot (N/D)$ , hence  $g_i = k_i \cdot (A/N) = d_i \cdot (N/D) \cdot (A/N) = d_i \cdot (A/D)$ , i.e., that indeed  $g_i = d_i \cdot (A/D)$ .

We have proved the desired equality (2) for all the cases when for all the debts  $d_i$ , we have  $d_i = k_i \cdot (D/N)$  for some integer  $k_i$ , i.e., when  $d_i/D = k_i/N$ . Any real number  $d_i/D$  can be approximated – with accuracy  $1/N$  – by an appropriate fraction  $k_i/N$ . As  $N$  increases, the fraction tends to  $d_i/D$ . Thus, since the solution  $S$  is continuous, in the limit, we will have (2) for all possible real values  $d_i$ .

The proposition is proven.

*Comment.* At first glance, it may sound reasonable to also require that if we combine two bankruptcy problems together, then in the combined problem, each creditors should receive the sum of what he/she would receive in each solutions. In other words:

- if we have  $S(\langle d_1, \dots, d_n \rangle) = \langle g_1, \dots, g_n \rangle$  and  $S(\langle d'_1, \dots, d'_n \rangle) = \langle g'_1, \dots, g'_n \rangle$ ,
- then we should have  $S(\langle d_1 + d'_1, \dots, d_n + d'_n \rangle) = \langle g_1 + g'_1, \dots, g_n + g'_n \rangle$ .

This requirement is explicitly mentioned in [7]. Let us show, however, that the fair solution does not have this property. Indeed:

- let us take  $d_1 = 4$ ,  $d_2 = 1$ , and  $A = 2$ , then  $D = d_1 + d_2 = 4 + 1 = 5$ , so  $A/D = 2/5 = 0.4$ ,  $g_1 = d_1 \cdot (A/D) = 4 \cdot 0.4 = 1.6$ , and  $g_2 = d_2 \cdot (A/D) = 1 \cdot 0.4 = 0.4$ ;
- let us also take  $d'_1 = d'_2 = 1$  and  $A' = 1$ , then  $D' = d'_1 + d'_2 = 1 + 1 = 2$ , so  $A'/D' = 1/2 = 0.5$ , and  $g'_i = d'_i \cdot (A'/D') = 1 \cdot 0.5 = 0.5$ .

On the other hand, for  $d_1 + d'_1 = 5$ ,  $d_2 + d'_2 = 2$ , and  $A + A' = 3$ , we have  $D + D' = 7$ , so  $(A + A')/(D + D') = 3/7$ . Thus, for the first creditor, the payment is

$$(d_1 + d'_1) \cdot ((A + A')/(D + D')) = 5 \cdot (3/7) = 15/7 = 2 + 1/7,$$

which is different from this creditor's summary payment  $g_1 + g'_1 = 1.6 + 0.5 = 2.1$  in two original situations.

## 4 Case of Interval Uncertainty

**Interval Sum and Interval Order: Reminder.** In the case of interval uncertainty, if we only know that the debt  $d_1$  is in the interval  $[\underline{d}_1, \bar{d}_1]$  and that the debt  $d_2$  is in the interval  $[\underline{d}_2, \bar{d}_2]$ , then the only conclusion we can make about the summary debt  $d_1 + d_2$  to these two creditors is that this sum belongs to the interval

$$[\underline{d}_1 + \underline{d}_2, \bar{d}_1 + \bar{d}_2].$$

This interval is known as the *sum*  $[\underline{d}_1, \bar{d}_1] + [\underline{d}_2, \bar{d}_2]$  of the two intervals  $[\underline{d}_1, \bar{d}_1]$  and  $[\underline{d}_2, \bar{d}_2]$ ; see, e.g., [4, 9, 11].

A natural order is component-wise: we say that the debt  $[\underline{d}_i, \bar{d}_i]$  to creditor  $i$  is smaller than or equal to the debt  $[\underline{d}_j, \bar{d}_j]$  to creditor  $j$  if  $\underline{d}_i \leq \underline{d}_j$  and  $\bar{d}_i \leq \bar{d}_j$ .

**Definition 4.** Let  $A$  be a positive real numbers and let  $[\underline{D}, \bar{D}]$  be an interval for which  $0 < \underline{D}$  and  $A < \bar{D}$ .

- We will call  $A$  the amount of assets, and we will call  $[\underline{D}, \bar{D}]$  the amount of debt.
- By a solution to the bankruptcy problem (or simply solution, for short), we mean a function  $S$  that maps every tuple  $\langle [\underline{d}_1, \bar{d}_1], \dots, [\underline{d}_n, \bar{d}_n] \rangle$  of intervals for which  $0 \leq \underline{d}_i$ , numbers for which  $\underline{d}_1 + \dots + \underline{d}_n = \underline{D}$ , and  $\bar{d}_1 + \dots + \bar{d}_n = \bar{D}$  into the same-size tuple of non-negative real numbers  $\langle g_1, \dots, g_n \rangle$  for which

$$g_1 + \dots + g_n = A.$$

**Definition 5.** We say that the solution  $S$  is fair if the following two requirements are satisfied when  $S(\langle [\underline{d}_1, \bar{d}_1], \dots, [\underline{d}_n, \bar{d}_n] \rangle) = \langle g_1, \dots, g_n \rangle$ :

- if  $\underline{d}_i \leq \underline{d}_j$  and  $\bar{d}_i \leq \bar{d}_j$ , then  $g_i \leq g_j$ ;
- $S(\langle [\underline{d}_1 + \underline{d}_2, \bar{d}_1 + \bar{d}_2], [\underline{d}_3, \bar{d}_3], \dots, [\underline{d}_n, \bar{d}_n] \rangle) = \langle g_1 + g_2, g_3, \dots, g_n \rangle$ .

**Definition 6.** We say that the solution  $S$  is continuous, if for every  $n$ , if  $\underline{d}_i^{(k)} \rightarrow \underline{d}_i$  and  $\overline{d}_i^{(k)} \rightarrow \overline{d}_i$  for all  $i$ ,  $S(\langle [\underline{d}_1^{(k)}, \overline{d}_1^{(k)}], \dots, [\underline{d}_n^{(k)}, \overline{d}_n^{(k)}] \rangle) = \langle g_1^{(k)}, \dots, g_n^{(k)} \rangle$ , and  $g_i^{(k)} \rightarrow g_i$  for all  $i$ , then

$$S(\langle [\underline{d}_1, \overline{d}_1], \dots, [\underline{d}_n, \overline{d}_n] \rangle) = \langle g_1, \dots, g_n \rangle.$$

**Proposition 2.** For each solution  $S$ , the following two conditions are equivalent to each other:

- the solution is fair and continuous,
- for some  $\alpha \in [0, 1]$ , the solution has the form  $g_i = d_i \cdot (A/D)$ , where

$$d_i \stackrel{\text{def}}{=} \alpha \cdot \overline{d}_i + (1 - \alpha) \cdot \underline{d}_i \text{ and } D \stackrel{\text{def}}{=} \alpha \cdot \overline{D} + (1 - \alpha) \cdot \underline{D}.$$

*Comment.* So, the solutions based on Hurwicz combinations  $d_i = \alpha \cdot \overline{d}_i + (1 - \alpha) \cdot \underline{d}_i$  are the only one which are fair (and continuous).

**Proof.** It is east to check that the solution based on Hurwicz combination is fair and continuous. So, to complete the proof, it is sufficient to prove that every fair continuous solution  $S$  has this form.

Indeed, let  $S$  be a fair and continuous solution. For every natural number  $N$ , we can consider the tuple

$$\langle [\underline{D}/N, \underline{D}/N], \dots, [\underline{D}/N, \underline{D}/N], [0, (\overline{D} - \underline{D})/N], \dots, [0, (\overline{D} - \underline{D})/N] \rangle \tag{3}$$

consisting of:

- $N$  degenerate debt intervals  $[\underline{D}/N, \underline{D}/N]$  and
- $N$  intervals  $[0, (\overline{D} - \underline{D})/N]$ .

By the first fairness requirement, since the debts  $d_i$  are the same for all first  $N$  creditors, the payments  $g_i$  should also be all equal  $g_1 = \dots = g_N$ . Similarly, the payments to the last  $N$  creditors should be the same:  $g_{N+1} = \dots = g_{2N}$ .

For any two sequences of natural numbers  $k_1, \dots, k_n, \ell_1, \dots, \ell_n$  for which

$$k_1 + \dots + k_n = \ell_1 + \dots + \ell_n = N,$$

the tuple

$$\langle [k_1 \cdot (\underline{D}/N), k_1 \cdot (\underline{D}/N) + \ell_1 \cdot (\overline{D} - \underline{D})/N], \dots, [k_n \cdot (\underline{D}/N), k_n \cdot (\underline{D}/N) + \ell_n \cdot (\overline{D} - \underline{D})/N] \rangle$$

can be obtained from the tuple (3) by adding up:

- the first  $k_1$  intervals from the first half and the first  $\ell_1$  intervals from the second half, then
- the next  $k_2$  intervals from the first half and the next  $\ell_2$  intervals from the second half, etc.



So, due to the second fairness requirements, the resulting payment tuple  $\langle g_1, \dots, g_n \rangle$  can be obtained from the tuple  $\langle g_1, \dots, g_1, g_{N+1}, \dots, g_{N+1} \rangle$  by adding the corresponding payment terms. Thus, the payment tuple has the form

$$\langle k_1 \cdot g_1 + \ell_1 \cdot g_{N+1}, \dots, k_n \cdot g_1 + \ell_n \cdot g_{N+1} \rangle.$$

In other words, for each debt interval

$$[\underline{d}_i, \bar{d}_i] = [k_i \cdot (\underline{D}/N), k_i \cdot (\underline{D}/N) + \ell_i \cdot (\bar{D} - \underline{D})/N],$$

the payment is equal to

$$g_i = k_i \cdot g_1 + \ell_i \cdot g_{N+1}. \tag{4}$$

Here,  $\underline{d}_i = k_i \cdot (\underline{D}/N)$ , so  $k_i = \underline{d}_i \cdot (N/\underline{D})$ . Similarly,  $\bar{d}_i - \underline{d}_i = \ell_i \cdot ((\bar{D} - \underline{D})/N)$  so  $\ell_i = (\bar{d}_i - \underline{d}_i) \cdot (N/(\bar{D} - \underline{D}))$ . Substituting these expressions for  $k_i$  and  $\ell_i$  into the formula (3), we conclude that  $g_i = a \cdot \underline{d}_i + b \cdot (\bar{d}_i - \underline{d}_i)$ , where we denoted  $a \stackrel{\text{def}}{=} g_1 \cdot (N/\underline{D})$  and  $b \stackrel{\text{def}}{=} g_{N+1} \cdot (N/(\bar{D} - \underline{D}))$ . Thus, we have

$$g_i = b \cdot \bar{d}_i + (a - b) \cdot \underline{d}_i. \tag{5}$$

The first fairness requirement means that if  $\bar{d}_i$  is larger then  $\bar{d}_j$  while  $\underline{d}_i = \underline{d}_j$ , then  $g_i$  should be larger (or the same) than  $g_j$ . This implies that  $a \geq 0$ . Similarly, if  $\underline{d}_i$  is larger then  $\underline{d}_j$  while  $\bar{d}_i = \bar{d}_j$ , then  $g_i$  should be larger (or the same) than  $g_j$ . This implies that  $a - b \geq 0$ .

Let us denote the ratio  $b/a$  by  $\alpha$ . Then,  $b = a \cdot \alpha$  and  $a - b = a \cdot (1 - \alpha)$ . Thus, the formula (5) takes the form

$$g_i = a \cdot (\alpha \cdot \bar{d}_i + (1 - \alpha) \cdot \underline{d}_i). \tag{6}$$

The sum of all the payments is equal to  $A$ , so

$$g_1 + \dots + g_n = a \cdot (\alpha \cdot \bar{d}_1 + (1 - \alpha) \cdot \underline{d}_1 + \dots + \alpha \cdot \bar{d}_n + (1 - \alpha) \cdot \underline{d}_n) =$$

$$a \cdot (\alpha \cdot (\bar{d}_1 + \dots + \bar{d}_n) + (1 - \alpha) \cdot (\underline{d}_1 + \dots + \underline{d}_n)) = a \cdot (\alpha \cdot \bar{D} + (1 - \alpha) \cdot \underline{D}) = a \cdot D,$$

hence  $a = A/D$  and the formula (6) takes the desired form

$$g_i = (\alpha \cdot \bar{d}_i + (1 - \alpha) \cdot \underline{d}_i) \cdot (A/D). \tag{7}$$

We have proved the desired equality (7) for all the cases when for all the creditors  $i$ , we have  $\underline{d}_i = k_i \cdot (\underline{D}/N)$  for some integer  $k_i$  and  $\bar{d}_i - \underline{d}_i = \ell_i \cdot ((\bar{D} - \underline{D})/N)$  for some integer  $\ell_i$ . Similarly to the proof of Proposition 1, any two real numbers can be thus approximated, and the larger  $N$ , the more accurate the resulting approximation. Thus, due to continuity, in the limit  $N \rightarrow \infty$ , we have (7) for all possible values  $\underline{d}_i$  and  $\bar{d}_i$ .

The proposition is proven.

**First Comment: What if We have Fuzzy Uncertainty?** For each creditor, instead of a single interval, we can have different intervals  $[\underline{d}_i(\alpha), \bar{d}_i(\alpha)]$  containing  $d_i$  with different degrees of uncertainty  $\alpha \in [0, 1]$ . If we pick a narrower sub-interval, then we become

less certain that  $d_i$  belongs to this sub-interval than that it belongs to the original interval. Thus, the interval corresponding to a higher degree of uncertainty is a subset of the interval corresponding to a lower degree of uncertainty. Such a sequence of embedded intervals is, in effect, an equivalent representation of a so-called *fuzzy number* (see, e.g., [1, 5, 10, 12, 13, 16]) for which the corresponding intervals are known as  $\alpha$ -cuts.

In this case, to describe each creditor’s debt, instead of two values  $\underline{d}_i$  and  $\bar{d}_i$ , we need to describe infinitely many values  $\underline{d}_i(\alpha)$  and  $\bar{d}_i(\alpha)$  corresponding to different  $\alpha \in [0, 1]$ . The overall debt corresponding to different  $\alpha$  can be obtained by adding all  $n$  debts:  $\underline{D}(\alpha) = \underline{d}_1(\alpha) + \dots + \underline{d}_n(\alpha)$  and  $\bar{D}(\alpha) = \bar{d}_1(\alpha) + \dots + \bar{d}_n(\alpha)$ .

Arguments similar to the ones we used in the proof of Proposition 2 lead to a conclusion that a fair solution is proportional to the linear combination  $d_i$  of these values, i.e., has the form  $g_i = d_i/D$ , where

$$d_i \stackrel{\text{def}}{=} \int (f_-(\alpha) \cdot \underline{d}_i(\alpha) + f_+(\alpha) \cdot \bar{d}_i(\alpha)) d\alpha$$

for some functions (maybe generalized functions)  $f_{\pm}(\alpha)$ , and

$$D \stackrel{\text{def}}{=} \int (f_-(\alpha) \cdot \underline{D}(\alpha) + f_+(\alpha) \cdot \bar{D}(\alpha)) d\alpha.$$

**What if We have Probabilistic Uncertainty?** What if for each  $d_i$ , we only know the probability distribution? In this case, it makes sense to use the following additional requirement on the bankruptcy solutions: that if we repeat the same division situation several ( $N$ ) times, the payments in the resulting overall situation should be  $N$  times larger. In the overall situation, the debt amount  $D_i$  is equal to the sum of  $N$  independent equally distributed debt amounts:  $D_i = d_i^{(1)} + \dots + d_i^{(N)}$ . According to the Large Numbers Theorem (see, e.g., [14]), for large  $N$ , the average

$$\frac{D_i}{N} = \frac{d_i^{(1)} + \dots + d_i^{(N)}}{N}$$

tends to the mean  $E[d_i]$  as  $N$  increases. Thus, for large  $N$ , the sum is getting (relatively) closer and closer to a single value –  $N$  times the mean. So, for large  $N$ , we have, in effect, the division problem in which instead of the original random variables, we have  $N$  times their means. The payments in the original problem should be  $N$  times smaller, i.e., they should be simply equal to the division corresponding to the means.

Thus, in the probabilistic case, we should simply compute the mean values  $E[d_i]$  of the debt amount, and distribute the assets proportionally to these mean values:

$$g_i = \frac{E[d_i]}{E[d_1] + \dots + E[d_n]} \cdot A.$$

**Acknowledgments.** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Belohlavek, R., Dauben, J.W., Klir, G.J.: *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford University Press, New York (2017)
2. Branzei, R., Dimitrov, D., Pickl, S., Tijs, S.: How to cope with division problems under interval uncertainty of claims? *Int. J. Uncertain. Fuzziness* **12**, 191–200 (2004)
3. Hurwicz, L.: *Optimality Criteria for Decision Making Under Ignorance*, Cowles Commission Discussion Paper, Statistics, No. 370 (1951)
4. Jaulin, L., Kiefer, M., Didrit, O., Walter, E.: *Applied Interval Analysis, With Examples in Parameter and State Estimation, Robust Control, and Robotics*. Springer, London (2001). <https://doi.org/10.1007/978-1-4471-0249-6>
5. Klir, G., Yuan, B.: *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, Upper Saddle River, New Jersey (1995)
6. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) *Human-Centric Decision-Making Models for Social Sciences*. SCI, vol. 502, pp. 163–193. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-642-39307-5\\_8](https://doi.org/10.1007/978-3-642-39307-5_8)
7. Li, X., Li, Y., Zheng, W.: Division schemes under uncertainty of claims. *Kybernetika* **57**(5), 849–855 (2021)
8. Luce, R.D., Raiffa, R.: *Games and Decisions: Introduction and Critical Survey*. Dover, New York (1989)
9. Mayer, G.: *Interval Analysis and Automatic Result Verification*. de Gruyter, Berlin (2017)
10. Mendel, J.M.: *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*. Springer, Cham, Switzerland (2017). <https://doi.org/10.1007/978-3-319-51370-6>
11. Moore, R.E., Kearfott, R.B., Cloud, M.J.: *Introduction to Interval Analysis*. SIAM, Philadelphia (2009)
12. Nguyen, H.T., Walker, C.L., Walker, E.A.: *A First Course in Fuzzy Logic*. Chapman and Hall/CRC, Boca Raton, Florida (2019)
13. Novák, V., Perfilieva, I., Močkoř, J.: *Mathematical Principles of Fuzzy Logic*. Kluwer, Boston, Dordrecht (1999)
14. Sheskin, D.J.: *Handbook of Parametric and Non-Parametric Statistical Procedures*. Chapman & Hall/CRC, London, UK (2011)
15. Yager, R.R., Kreinovich, V.: Fair division under interval uncertainty. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst. (IJUFKS)* **8**(5), 611–618 (2000)
16. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)



# Economy-Related Emotional Attitudes Towards Other People: How Can We Explain Them?

Christopher Reyes<sup>1</sup>, Vladik Kreinovich<sup>1(✉)</sup>, and Chon Van Le<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Texas at El Paso, 500 W. University,  
El Paso, TX 79968, USA

creyes24@miners.utep.edu, vladik@utep.edu

<sup>2</sup> International University of Ho-Chi-Minh City, Ho Chi Minh City, Vietnam  
lvchon@hcmiu.edu.vn

**Abstract.** Research has shown that to properly understand people's economic behavior, it is important to take into account their emotional attitudes towards each other. Behavioral economics shows that different attitudes results in different economy-related behavior. A natural question is: where do these emotional attitudes come from? We show that, in principle, such emotions can be explained by people's objective functions. Specifically, we show it on the example of a person whose main objective is to increase his/her country's GDP: in this case, the corresponding optimization problem leads exactly to natural emotions towards people who contribute a lot or a little towards this objective.

## 1 Formulation of the Problem

**Economy-Related Emotions are Important.** Traditional economics considers people as rational decision makers, that make all investment and other economic decisions based on the cold calculations of possible benefits and drawbacks of different options. In reality, people often have strong economy-related emotions, and these emotions affect human decisions. It is therefore important to take these emotions into account when predicting how people will behave.

Taking such emotions into account is an important part of behavioral economics, a branch of economics that recently got several Nobel prizes.

**But where do these Emotions Come From?** A natural next question is: where do these emotions come from? These emotions affect how people make economic decisions and thus, affect the country's economy. So, if a person wants the country's economy to be going in a certain direction, a natural hypothesis is that this person's economy-related emotions should help drive the country's economy in this direction.

In this paper, we show that this hypothesis indeed explains – at least on the qualitative level – people's economy-related emotions. We show it on the example of a situation when a person is mostly interested in increasing the country's Gross Domestic Product (GDP) as much as possible. We show that in this case, the analysis of the corresponding optimization problem leads exactly to the economy-related emotional attitudes that people experience.

## 2 Towards Formulating the Problem in Precise Terms

**How Decision Theory Describes Individual and Group Decision Making.** According to decision theory (see, e.g., [2–5, 8–10]), decisions of a rational person  $i$ , i.e., a person whose decisions are consistent, are equivalent to optimizing an appropriate function  $u_i(x)$  known as *utility function*. In other words, decisions of a rational person  $i$  are equivalent to selecting an alternative  $x$  for which the utility  $u_i(x)$  is the largest possible.

In general, utility is defined modulo linear transformations: instead of the original function  $u_i(x)$ , we can use an alternative function  $u'_i(x) = a \cdot u_i(x) + b_i$  for some constants  $a_i > 0$  and  $b_i$ ; this new function describes exactly the same preferences and thus, exactly the same economic behavior.

What if several people need to make a joint decision affecting all of them? In this paper, we consider a what is called a *win-win* situation, when we need to select between several decisions each of which is potentially beneficial for everyone. In such cases, we start with what is called a *status quo* situation  $x_0$  – the situation in which the group is right now (and in which the group will remain if no group decision is selected). In this case, to make analysis easier, it makes sense to re-scale all individual utilities so that each person utility  $u_i(x_0)$  of the status quo situation  $x_0$  becomes 0. This can be done, e.g., by going from the original scale  $u_i(x)$  to the new scale  $u_i(x) + b_i$  with  $b_i = -u_i(x_0)$ . Because of this possibility, in the following text, we will assume that all utility functions already have this property, i.e., that  $u_i(x_0) = 0$  for all participants  $i$ .

Under this assumption, decision theory recommends to select a decision  $x$  for which the product of the utilities  $\prod_{i=1}^n u_i(x)$  is the largest possible. This idea is known as *Nash's bargaining solution*; see, e.g., [5–7].

**How Emotional Attitudes Towards other People are Taken into Account.** Emotional attitude means that the person's preferences – and thus, the person's utility function  $u_i(x)$  that describes these preferences – are affected not only by the objective conditions of this person, but also by the conditions (i.e., utilities) of others. Let us denote the utility that only takes into account the objective conditions by  $u_i^{(0)}(x)$ . The actual utility  $u_i(x)$  is affected not only by this value  $u_i^{(0)}(x)$ , but also by utilities  $u_j(x)$  of others:

$$u_i(x) = f_i(u_i^{(0)}(x), u_j(x), u_{j'}(x), \dots).$$

The effect of others is usually smaller than the effect of the person's own objective conditions. Since the effect of the values  $u_j(x)$  is small, we can follow the usual practice of physics and other applications (see, e.g., [1, 11]): expand the dependence on these values in Taylor series and keep only linear terms in this expansion. So, we end up with the following formula:

$$u_i(x) = u_i^{(0)}(x) + \sum_{j \neq i} \alpha_{ij} \cdot u_j(x), \tag{1}$$

for appropriate coefficients  $\alpha_{ij}$ . These coefficients  $\alpha_{ij}$ , in effect, describe the emotions of the  $i$ -th person toward a person  $j$ :

- When the coefficient  $\alpha_{ij}$  is positive, this means positive attitude: the person  $i$  feels better when he/she knows that the person  $j$  is better.

- When the coefficient  $\alpha_{ij}$  is negative, this means negative attitude: the more person  $j$  enjoys life, the worse person  $i$  feels. This negative feeling may be well-justified: e.g., when the person  $j$  gained his money in a still-legal but highly unethical way, by hurting others.

**Resulting Formulation of the Problem.** Suppose that a person  $i$  wants the community to achieve a certain objective – e.g., to increase the overall GDP which can be approximately described as the sum

$$G \stackrel{\text{def}}{=} u_i^{(0)} + \sum_{j \neq i} u_j. \tag{2}$$

The person  $i$  can change the group behavior by using appropriate emotions toward other people. Indeed, once the person  $i$  fixes his/her emotions, i.e., the coefficients  $\alpha_{ij}$ , then, according to the Nash’s bargaining solution, the group will select the alternative that maximizes the product

$$F \stackrel{\text{def}}{=} \left( u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j \right) \cdot \prod_{j \neq i} u_j. \tag{3}$$

The question is: what coefficients  $\alpha_{ij}$  should the person  $i$  select so that the result of maximizing the expression (4) will also maximize  $i$ -th objective  $G$  – e.g., in our example, the expression (2).

### 3 Analysis of the Problem

Let us first formulate the above problem in general mathematical terms. We have two functions  $F(v_1, \dots, v_n)$  and  $G(v_1, \dots, v_n)$  of several variables. We want to make sure that at the point  $m = (m_1, \dots, m_n)$  at which the first function attains its maximum under some constraints, the second function also attains its largest value under the same constraints.

The fact that at the point  $m$ , the function  $F(v_1, \dots, v_n)$  attains its maximum under give constraints means that for any perturbation  $m_i \mapsto m_i + \Delta m_i$  which is consistent with these constraints, the value of this function cannot increase. In particular, this must be true for small perturbations  $\Delta m_i$ . For small perturbations, terms quadratic (and of higher order) with respect to these perturbations are very small and can, thus, be safely ignored. Thus, to find the modified value  $F(m_1 + \Delta m_1, \dots, m_n + \Delta m_n)$  of this function, we can expand this expression in Taylor series in terms of  $\Delta m_i$  and keep only linear terms in this expansion. In this case, we get

$$F(m_1 + \Delta m_1, \dots, m_n + \Delta m_n) = F(m_1, \dots, m_n) + \sum_{i=1}^n \frac{\partial F}{\partial m_i} \cdot \Delta m_i. \tag{4}$$

Thus, the requirement that the value of the function  $F(v_1, \dots, v_n)$  attains its maximum means that for all possible perturbations  $\Delta m_i$ , the new value

$$F(m_1 + \Delta m_1, \dots, m_n + \Delta m_n)$$

of this function is smaller than or equal to the previous value  $F(m_1, \dots, m_n)$ . Due to the formula (4), this difference is equal to the sum in the right-hand side of this formula. Thus, the maximizing condition means that this sum should be non-positive:

$$\sum_{i=1}^n \frac{\partial F}{\partial m_i} \cdot \Delta m_i \leq 0. \quad (5)$$

This sum is the scalar (“dot”) product  $\nabla F \cdot \Delta m$  of two vectors: the gradient vector

$$\nabla F = \left( \frac{\partial F}{\partial m_1}, \dots, \frac{\partial F}{\partial m_n} \right) \quad (6)$$

and the perturbations vector

$$\Delta m = (\Delta m_1, \dots, \Delta m_n). \quad (7)$$

Thus, the fact that the function  $F(v_1, \dots, v_n)$  attains its maximum at the point  $m$  implies that for all possible perturbations  $\Delta m$ , we have  $\nabla F \cdot \Delta m \leq 0$ .

The fact that at the same point  $m$ , the function  $G$  should not increase means that  $\Delta G \cdot \Delta m \leq 0$ . We do not exactly know a priori which perturbations  $\Delta m$  will be possible and which not. So, to make sure that the maximum of  $F$  also implies the maximum of  $G$ , it is reasonable to require that for *all* possible vectors  $\Delta m$ , if we have  $\nabla F \cdot \Delta m \leq 0$ , then we should also have  $\nabla G \cdot \Delta m \leq 0$ .

In particular, if  $\nabla F \cdot \Delta m = 0$ , this means that we have both  $\nabla F \cdot \Delta m \leq 0$  and  $\nabla F \cdot (-\Delta m) \leq 0$ . Thus, we should have  $\nabla G \cdot \Delta m \leq 0$  and  $\nabla G \cdot (-\Delta m) \leq 0$  – i.e.,  $\nabla G \cdot \Delta m \geq 0$ . So, we should have  $\nabla G \cdot \Delta m = 0$ . In geometric terms, the fact that the dot product of two vectors is 0 means that these vectors are orthogonal to each other. Thus, every vector  $\Delta m$  which is orthogonal to  $\nabla F$  should be orthogonal to  $\nabla G$ . All the vectors orthogonal to a given vector  $\nabla F$  form a (hyper-)plane orthogonal to this vector. It is known that all the vectors which are orthogonal to all the vectors from this plane are collinear with  $\nabla F$ , i.e., we must have  $\nabla G = c \cdot \nabla F$  for some constant  $c$  – or, equivalently, that  $\nabla F = c' \cdot \nabla G$  for some constant  $c' = 1/c$ .

Let us use this conclusion to analyze our case study, in which we unknowns  $v_i$  are:

- the “objective” utility value  $u_i^{(0)}$  of person  $i$ , and
- the utility values  $u_j$  corresponding to all other persons  $j$ .

## 4 Case Study

**Description of the Case: Reminder.** We consider the case when the main objective of the person  $i$  is increasing the GDP of his/her country.

In this case, the function  $G$  has the form (2).

**Analysis of the Case.** For the function  $G$ , its gradient is equal to  $\nabla G = (1, \dots, 1)$ , so the above condition means that

$$\nabla F = c' \cdot \nabla G = (c', \dots, c') \quad (8)$$

for some constant  $c'$ , i.e., that all partial derivatives of the function  $F$  have the same value. It is convenient to describe  $F$  as  $F = \exp(H)$ , where

$$H = \ln(F) = \ln \left( u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j \right) + \sum_{j \neq i} \ln(u_j). \quad (9)$$

Here, by the chain rule formula,  $\nabla F = \exp(H) \cdot \nabla H$ . So, all components of the vector  $\nabla H$  differ from the corresponding components of the vector  $\nabla F$  by the same factor  $F = \exp(H)$ . Since all the components of the gradient  $\nabla F$  are equal to each other, this implies that all the components of the gradient  $\nabla H$  are also equal to each other.

Differentiating the expression (9) with respect to  $u_i^{(0)}$ , we conclude that

$$H_{,i} \stackrel{\text{def}}{=} \frac{\partial H}{\partial u_i^{(0)}} = \frac{1}{u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j}. \quad (10)$$

For each  $k \neq i$ , differentiating the expression (9) with respect to  $u_k$ , we get:

$$H_{,k} \stackrel{\text{def}}{=} \frac{\partial H}{\partial u_k} = \frac{\alpha_{ik}}{u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j} + \frac{1}{u_k}. \quad (11)$$

These two derivative – i.e., these two components of the gradient – must be equal to each other, i.e., we must have

$$\frac{\alpha_{ik}}{u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j} + \frac{1}{u_k} = \frac{1}{u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j}. \quad (12)$$

Multiplying both sides of this equation by

$$C \stackrel{\text{def}}{=} u_i^{(0)} + \sum_{j \neq i} \alpha_{ij} \cdot u_j, \quad (13)$$

we conclude that

$$\alpha_{ik} + \frac{C}{u_k} = 1. \quad (14)$$

Thus, we arrive at the following formula for the coefficients  $\alpha_{ik}$  describing the  $i$ -th person's emotions towards others.

**Resulting Formula and Its Interpretation.** For a person  $i$  whose main objective is increasing the country's GDP, the appropriate emotions towards others – namely, the emotions that best promote this objective – are described by the formula

$$\alpha_{ik} = 1 - \frac{C}{u_k}. \quad (15)$$



Thus:

- When a person  $k$  works hard and contributes a lot to the GDP – and thus, get a lot of compensation  $u_k$  for his/her hard work, we get  $\alpha_{ik} \approx 1$  – i.e., the person  $i$  has a very positive attitude towards this hard-working person  $k$ .
- On the other hand, if a person  $k$  works as little as possible, so that  $k$ 's compensation is small, the  $i$ 's attitude towards  $k$  is much less positive, and it can be even negative if  $u_k < C$ .

*Comments.*

- From the commonsense viewpoint, this negative attitude makes sense: if  $i$ 's goal is to increase the country's GDP, then  $i$  naturally feels negative towards those who could help their country more but prefer not to work too hard. What we showed is that not only such motions are natural, they actually help achieve such economic goals. For example, if many people think like that, the country may try to force people to work more – e.g., by imposing special taxes on those who do not pull their share of effort.
- It is important to take into account that we are dealing with an approximate model and thus, our main conclusion – the formula (15) – should not be taken too literally. For example, it is necessary to take into account that the formula (15) – and the resulting negative attitude – only make sense towards people who could work more but prefer not to. It does not make any economic sense to have negative feelings towards people who try their best but cannot produce too much because of their health or age or disability.

**Acknowledgments.** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Feynman, R., Leighton, R., Sands, M.: The Feynman Lectures on Physics. Addison Wesley, Boston, Massachusetts (2005)
2. Fishburn, P.C.: Utility Theory for Decision Making. Wiley, New York (1969)
3. Fishburn, P.C.: Nonlinear Preference and Utility Theory. The John Hopkins Press, Baltimore, Maryland (1988)
4. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) Human-Centric Decision-Making Models for Social Sciences. SCI, vol. 502, pp. 163–193. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-642-39307-5\\_8](https://doi.org/10.1007/978-3-642-39307-5_8)
5. Luce, R.D., Raiffa, R.: Games and Decisions: Introduction and Critical Survey. Dover, New York (1989)
6. Nash, J.: The bargaining problem. *Econometrica* **18**(2), 155–162 (1950)

7. Nguyen, H.P., Bokati, L., Kreinovich, V.: New (simplified) derivation of Nash's bargaining solution. *J. Adv. Comput. Intell. Intell. Inform. (JACIII)* **24**(5), 589–592 (2020)
8. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction. *Int. J. Intell. Syst.* **24**(1), 27–47 (2009)
9. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer Verlag, Berlin, Heidelberg (2012)
10. Raiffa, H.: *Decision Analysis*. McGraw-Hill, Columbus, Ohio (1997)
11. Thorne, K.S., Blandford, R.D.: *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton University Press, Princeton, New Jersey (2017)



# COVID-19 and Short-Run Survival in the Service Sector: Evidence from the Tourism Economy

Surapot Baiya<sup>1</sup>(✉), Pithoon Thanabordeekij<sup>1</sup>, and Paravee Maneejuk<sup>2</sup>

<sup>1</sup> Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand  
{surapot.b,pithoon.th}@cmu.ac.th

<sup>2</sup> Centre of Excellence in Econometrics, Faculty of Economics,  
Chiang Mai University, Chiang Mai, Thailand

**Abstract.** This study aims to evaluate the survival probability of small- and medium-sized enterprises (SMEs) in the service sector during the COVID-19 pandemic and empirically investigate their survival determinants. The sample group of 420 firms was analyzed utilizing the Cox proportional hazard model. The results showed that several business categories had various levels of survivability. The businesses that served exposure to tourists (e.g., travel agencies, entertainment, accommodation, and restaurant) were more likely to fail. Furthermore, business characteristics, financial statements, and government support substantially impacted the firm survival likelihood. These findings have several practical implications for businesses, governments, and policymakers in dealing with pandemics in the future.

**Keywords:** Firm survival · COVID-19 · Service sector · Financial ratio · Cox proportional hazard

## 1 Introduction

The world economy is currently facing a global economic crisis due to the COVID-19 pandemic, which has been declared a public health emergency of international concern (PHEIC) by the World Health Organization (WHO). Many countries have imposed limitations, such as social isolation and mask-wearing, to prevent the spread of COVID-19 [2]. The COVID-19 pandemic has interrupted the activities of countries' economies. As a result of COVID-19, small- and medium-sized enterprises (SMEs) are confronting a variety of concerns and problems, including the freezing of business activities, weakened financial situations, and exposure to financial risk [49,55]. Many small businesses cannot deal with these circumstances, resulting in their closure in the first months of the pandemic [10]. The changing environment has led companies to develop and seek essential resources to cope with the economic crisis.

Various studies relevant to the effects of pandemics on the service industry have reported that business related to tourism is highly vulnerable to epidemic outbreaks and threats [40, 50, 59]. COVID-19 is easy to transmit and causes international anxiety. As a result, traveling is restricted by public health measures from the government immediately. However, service sector businesses might have activities directly or indirectly related to tourism, which is divided into two sub-groups: mainly tourism and partly tourism [25]. Considering this fact, we can reasonably expect that the survivability of businesses whose activities are offered mostly to tourism is lower than partly tourism. Hence, this study aims to investigate the effects of COVID-19 on firm survival in the short run. Various businesses in the service sector are compared in terms of their survival probabilities. The determinants of firm survival are examined as well.

In the empirical research on the industrial organization (IO), several studies have analyzed the survival of firms in manufacturing. The subjects of research have understandably focused on the evolution of the industry caused by changes in technology and innovation [3, 5, 6], the impact of the life cycle of the organization [4], the role of location as well as economic agglomeration [27], and the worldwide financial crisis's impact [43]. Past research has investigated the elements influencing firm longevity by differentiating internal and external influences [41]. Several studies suggest that the resource-based view is essential when analyzing internal resources [29]. Financial variables predict the firm's survival [30]. However, the financial statement of small and medium companies may be highly deceptive. Financial fraud may be a reliable sign of severe financial issues leading to bankruptcy. Thus, it may be challenging to forecast organizations' financial problems using their financial documents. The financial indicators derived from Kumar and Ravi [37] have been concluded to evaluate financial conditions.

Additionally, the economic crisis due to the COVID-19 pandemic is unique because this crisis will be driven by non-economic factors [12]. Therefore, dealing with such a crisis requires unique monetary and fiscal policies. Azad et al. [9] found that fiscal policy was more active than monetary policy because the short-term stimulus provided by deficit spending helped increase economic activity. The such study agreed with the study of Şengel et al. [56], which reported on the effects in countries that provided credit and liquidity to support the service sector through tax breaks and low-interest rates. In Thailand, the government has announced many policies to support SMEs, such as tax reduction and deferrals, employment cost support, soft loan, suspension of interest rates, credit deferral, and debt rollover [57]. Thus, external subsidies may increase the probability of firm survival.

It must be stressed that most firm survival studies are conducted on manufacturing and lack empirical survival analysis in the service sector, particularly during the COVID-19 spread. To fill the research gap, this study surveys the data from Chiang Mai Province, Thailand, as this province is the leading tourism economy [16]. By doing this, we fill a research vacuum concerning identifying firm survival in the service sector during the pandemic. This study synthesizes the theoretical and empirical research to improve the model for analysis of firm

survival by employing (1) a variety of finance variables in the literature (profitability, liability, efficiency, and leverage), (2) firm characteristics which used indicators in other empirical investigations (firm size, age, legal form, and business category), and (3) government support to examine the effect of policy support on firm survival. Moreover, we offer a concept for evaluating the short-run survival of a firm based on the Cox proportional hazard model.

In essence, this work adds to the body of knowledge in the following ways. First, this study employs firm-level data to assess firm survival in the service sector. Second, the study examines different risks of failure among businesses in the service sector. Third, the paper sheds light on the crucial financial statements in firm survival, uncovering suggestive evidence that financial strategy management is critical for firms to survive through the economic crisis. The rest of this paper is organized as follows. Section 2 reviews the theoretical as well as empirical studies on firm survival. Section 3 discusses the sample selection together with the methodology. Section 4 presents the study's findings. Finally, Sect. 5 contains the conclusions.

## 2 Literature Review and Hypotheses

Considering previous research on industrial organization, the theoretical background of firm survival has been based on the resource-based presumption and the organization ecology [24]. Several studies have indicated the factors of firm survival and differentiated between internal firm-specific characteristics and external environmental factors [41]. Thus, we adopt an eclectic approach based on the theoretical framework and empirical studies.

### 2.1 Resource-Based Theory

The resource-based concept contends that strategic resources are required to promote the firm's capacity to build competitive advantage, and the resources connected with corporate development are necessary to be investigated for their functions in the firm's success and survival [51]. According to Hitt et al. [34], resources can be divided into tangible and intangible. Tangible resources such as processing equipment, manufacturing facilities as well as distribution centers are assets that can be seen and quantified. Meanwhile, intangible resources are assets that are firmly embedded in the firm's past (accruing through time) and are remarkably difficult for competitors to determine and replicate. However, we only consider tangible assets in the form of financial variables. Brown et al. [14] found that financial resources may be advantageous for a firm to seek equity investment to improve its position in various resource component markets during the crisis. Previous studies on determinants of bankruptcy have employed many financial variables to predict insolvency [37]. Firm-specific financial indicators have also been reported to be one of the most widely studied aspects of survival analysis in emerging Asian economies [58]. To assess financial components in our survival model, Clementi and Hopenhayn et al. [19] suggested considering

three aspects of financial conditions (profitability, leverage, and asset) from the balance sheet. First, we define profitability as net profit divided by total assets. In other indicators that determine firm performance, we employ operational efficiency as total asset turnover, defined as net sales divided by the total asset. Additionally, Chen and LEE [18] suggested that liquidity determined a firm's survival capability. Thus, we define liquidity as a current asset over the current liability. Finally, leverage is defined as total debt divided by total assets to assess the firm's total indebtedness. Firms with larger debt often have weaker balance sheets, making it more difficult to receive funds from external finance [13].

## 2.2 Organization Ecology

According to the firm's theoretical model and industry dynamics, the failure rate fluctuates with firm age [21, 23, 36]. Startup firms are often smaller than established ones, making them more vulnerable to environmental changes. The young firms might not survive if they do not achieve the appropriate initial endowments and/or cannot build the necessary competencies [24]. Over time, the firm must understand its capacity for doing business. This learning process could take years. Thus, it is reasonable to anticipate a substantially greater exit rate in the business of new firms than older firms participating in the same market. Accordingly, the exit rate is anticipated to decline as organizations age [17, 28, 44]. This concept is also known as the "liability of newness". However, organization ecologists assert that there can be many relationships between firm age and survival [24]. Many investigations have discovered an inverse U-shaped. It is associated between failure rate and age, meaning that a new firm can survive for a while with a low likelihood of failure by using the initial stock of endowments (bank loan, venture capital) [15, 26].

Consequently, the failure rate peaks when initial resources are depleted and declines as only the market's most competitive firms survive. Several studies have reported the "liability of adolescence". Baum [11] and Hannan [31] have found the "liability of obsolescence", which refers to the likelihood of leaving that rises with time-varying, implying that older firms are highly inertial and are more likely to grow worse at adapting to new competitive environments. Prior studies have reported that firm survival was related to the initial size of the firm [1, 8]. However, Mata et al. [45] argued that the current size of the firm was a stronger indicator of a firm's survival chance than the original size since the current size incorporated data on how a firm responded to market success over time. In this regard, a firm's ability to adapt to a changing competitive environment is critical in determining its survival [39]. Whether the proxy size measure is employed or not, several studies have indicated that the chance of a firm's survival positively relates to its size. Larger companies are less susceptible than smaller companies operating on a smaller scale because their output levels are more likely to be near the industry's minimum efficient scale [7]. Hence, this work hypothesizes that the current size of the firm is an adequate explanatory parameter.

Furthermore, numerous studies have investigated the differences in the firm’s legal structure as a critical factor in firm survival. Harhoff et al. [32] discovered that unlimited liability companies had a higher survival rate than limited ones. Such a study agreed with the study of Baumöhl et al. [11] which found that the legal form of business affected business survival. Unlimited liability companies outlived limited liability companies, and various legal forms had distinct effects on the likelihood of survival. This argument leads us to set legal forms in our estimation.

The risk of firm failure might depend on business activity. Reis et al. [53] pointed out that the firms related to tourism are classified by the intensity of tourism: Mainly Tourism and Partly Tourism. Firms in the former category offer activities primarily to tourists, such as travel agencies and accommodation; the latter group comprises businesses, including restaurants, bars, and transportation that serve locals and offer services to tourists. Hence, firms whose activities are catered mainly to tourism have a high risk of failure [40].

Previous research works also provided hypotheses relating to critical factors impacting firm survival. Table 1 summarizes all the survival factors considered in this work and the predicted impacts of each factor. These impacts are based on the results of previous empirical research.

**Table 1.** Summary of variables used in the empirical research

Category	Factor	Expected hypothesis on firm survival
Firm characteristic	Firm size	+
	Firm age	+
	Legal form	+
	Sector	–
Financial statement	Profitability	+
	Liquidity	+
	Efficiency	+
	Leverage	–
External subsidy	Government support	+

### 3 Sample and Methodology

#### 3.1 Data

This study was conducted using the data gathered from 420 service enterprises in Chiang Mai Province, Thailand from an annual survey funded by the Department of Business Development (DBD) and the Office of SMEs Promotion (OSMEP). The firms that satisfied a condition (operating at the end of 2019, before the COVID-19 pandemic) were considered. Among the 420 firms,

106 failed after the COVID-19 outbreak; thus, the failure rate was 25.2%. More details about the descriptive data are provided in Table 2.

A review of the literature reveals that firm failure can be measured in several ways, such as a firm exit from the market [5], and firm liquidation [54]. Therefore, to achieve the purpose of the study, we identify which condition contributes to short-run survival during the crisis period. Several research works suggested that SMEs required sufficient short-term liquidity for at least six months to avoid insolvency [22,46]. Hence, we measured the survival time by requesting the owner's subjective evaluation of the firm's survival. For firms that survived up to the end of the study period, the value of zero was used as a variable status throughout the study period. In contrast, the value of one was used for a firm that did not survive till the end of the study period.

The list of firm characteristics used includes firm size and age defined by the natural logarithm of the number of employees and the years of operation, respectively. Its square is also used to account for non-linearities. The study captured the importance of legal form with three categories: sole proprietorship, partnership, and company. The legal form categories were established as dummy variables. According to Harhoff et al. [32], the owners of limited liability corporations are not liable for the debts of their companies. Limited liability firms tend to fail more than unlimited liability firms since the unlimited liability firms' owners are liable for any losses with their wealth. Thus, this work employed sole proprietorship as a default legal form category.

Regarding sector, this study identified nine businesses in the service sector based on the definition of the Department of Business Development's Thailand Standard Industrial Classification (TSIC): construction, travel agency, personal care, logistics, passenger land transport, entertainment, accommodation, restaurant, and real estate. The business categories were measured as dummy variables. According to Reis et al. [53], construction is not categorized as a tourism-related business. Thus, this study employed construction as a default category.

In terms of financial statements, this research employed four widely used financial variables: the return on asset (ROA) as a proxy for profitability, the current ratio as a proxy for liquidity, the asset turnover as a proxy for efficiency, and the debt ratio as a proxy for leverage. The return on assets was determined by:  $[(\text{net profit}/\text{total asset}) \times 100]$ . The current ratio was computed by:  $[(\text{current asset}/\text{current liability})]$ . The asset turnover was calculated by:  $[(\text{net sale}/\text{total asset})]$ . The debt ratio was calculated by:  $[(\text{total debt}/\text{total asset})]$ .

Finally, the subsidy represented a key role in firm survival, as described in Sengel et al. [56]. This work found that the fiscal and monetary policies (e.g., tax breaks, travel cost subsidies, soft loans, and asset warehousing) supported firm survival. The government support is measured by policy inclusion using a five-point Likert-type scale ranging from 1 (weakly support) to 5 (strongly support).



**Table 2.** Definition and descriptive statistics of variables

Variable	Definition	Descriptive statistics		
		Mean	S.D	Median
<b>Firm size and age</b>				
Size	Number of labors	6.630	6.393	5.000
Age	Year of operation	8.250	6.069	6.000
lnSize	Natural logarithm of the Size variable	1.592	0.723	1.609
lnAge	Natural logarithm of the Age variable	1.933	0.555	1.792
Size <sup>2</sup>	Squared value of the Size variable	84.45	207.0	25.00
Age <sup>2</sup>	Squared value of the Age variable	104.8	217.1	36.00
<b>Legal forms</b>				
Sole proprietorship	Dummy variable for a sole proprietorship	0.388	0.488	0.000
Partnership	Dummy variable for partnership	0.254	0.436	0.000
Limited liability company	Dummy variable for company	0.254	0.480	0.000
<b>Business categories</b>				
Construction	Dummy variable for construction	0.281	0.450	0.000
Travel agency	Dummy variable for travel agency	0.066	0.250	0.000
Personal care	Dummy variable for personal care	0.045	0.208	0.000
Logistic	Dummy variable for logistic	0.047	0.213	0.000
Passenger land transport	Dummy variable for passenger land transport	0.021	0.145	0.000
Entertainment	Dummy variable for entertainment	0.028	0.167	0.000
Accommodation	Dummy variable for accommodation	0.133	0.340	0.000
Restaurant	Dummy variable for restaurant	0.121	0.327	0.000
Real estate	Dummy variable for real estate	0.254	0.436	0.000
<b>Financial statements</b>				
Profitability	Return on asset (%) <sup>a</sup>	-5.947	43.27	-0.035
Liquidity	Current ratio (x) <sup>b</sup>	56.32	141.8	7.700
Efficiency	Total asset turnover (x) <sup>c</sup>	0.804	1.428	0.360
Leverage	Debt ratio (x) <sup>d</sup>	0.571	1.767	0.135
<b>Subsidy</b>				
Government support	Five-point Likert-type scale	1.567	0.657	1.500

**Notes:** <sup>a</sup> Calculated by: (net profit/total asset) × 100.

<sup>b</sup> Calculated by: (current asset/current liability).

<sup>c</sup> Calculated by: (net sale/total asset).

<sup>d</sup> Calculated by: (total debt/total asset).

### 3.2 Methodology

The Cox proportional hazard model is applied in our study with the purpose to analyze the impacts of selected variables on survival of the firms [20]. In particular, this model enables us to investigate how certain conditions affect the likelihood of a specific event occurring. This model measures the incidence or hazard rate which can be generated by

$$h(t|x_{i1}, \dots, x_{in}) = h_0(t) \exp(\beta_1 x_{i1} + \dots + \beta_n x_{in}), \quad h_0(t) > 0 \quad (1)$$

where  $h_0(t)$  is the baseline hazard function, and reflects the underlying hazard for subjects with all covariates  $x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}$  equal to zero (reference group).  $\beta_1, \beta_2, \dots, \beta_n$  represent the estimated coefficients. Note that the baseline hazard function  $h_0(t)$  is assumed to be arbitrary and no distributional assumptions are required to carry out the estimation of  $\beta$  or  $h_0(t)$ . Thus, the model is often referred to as semiparametric [38].

The hazard function in Eq. (1) is the rate of failure of a firm, given that the failure has not occurred prior to time  $t$ . The function  $h_0(t)$  shows how the hazard function changes with survival time. The other function  $\exp(\beta_1 x_{i1} + \dots + \beta_n x_{in})$  shows how the hazard function changes as a function of the subject covariates. If  $h(t | x_{i1} = \dots = x_{in} = 0)$  the hazard for the control group and  $h(t | x_{i1} = \dots = x_{in} = 1)$  the treated group is given, then the hazard ratio for these two observations becomes

$$\frac{h(t | x_i = 1)}{h(t | x_i = 0)} = \frac{h_0(t) \cdot \exp \left[ \sum_{i=1}^n \beta_i x_{in} = 1 \right]}{h_0(t) \cdot \exp \left[ \sum_{i=1}^n \beta_i x_{jn} = 0 \right]} = \exp [\beta_i], i = 1, \dots, n. \quad (2)$$

We can see that the hazards of the two groups remain proportional over time and  $\exp[\beta_i]$  is called the hazard ratio. To estimate Eq. (1), the logarithmic transformation is used to obtain

$$\log (h(t | x_{i1}, \dots, x_{in})) = \log (h_0(t)) + \beta_1 x_{i1} + \dots + \beta_n x_{in} \quad (3)$$

where  $\beta_1, \beta_2, \dots, \beta_n$  become log hazard ratio. We also note that the hazard ratio specifically displays the probability of a firm's failure when a specific covariate  $x$  shifts by 1 unit. A determinant (covariate  $x$ ) might be considered as a factor driving the firm failure if estimation is  $\exp [\beta_i] > 1$  and a preventive factor impeding a firm's failure if  $\exp [\beta_i] < 1$  [52].

## 4 Results

Table 3 reports the survival status of 420 firms in the service sector after COVID-19. The failure rate represented the proportion of firms that probably failed by six months after the pandemic. When the data were censored, the simple failure rate might not accurately reflect the real risk of failure. Consequently, we utilized the Nelson-Aalen estimate of cumulative hazard estimator, which was tailored to data subject to right censoring. Based on Table 3, 106 out of 420 firms (accounted for 25.2%) failed during the duration of the study. The cumulative hazard function for the service industry estimated by the Nelson-Aalen method was found to have a coefficient value of 0.279. This result demonstrated the remarkable differences between business categories from this perspective: the failure rate for accommodation showed the highest failure rate at 0.554 (0.725), while construction had the lowest value of 0.102 (0.105). The risk in the service industries depended on the types of business activities. Firms that served tourists typically faced higher risks than those that served residents.

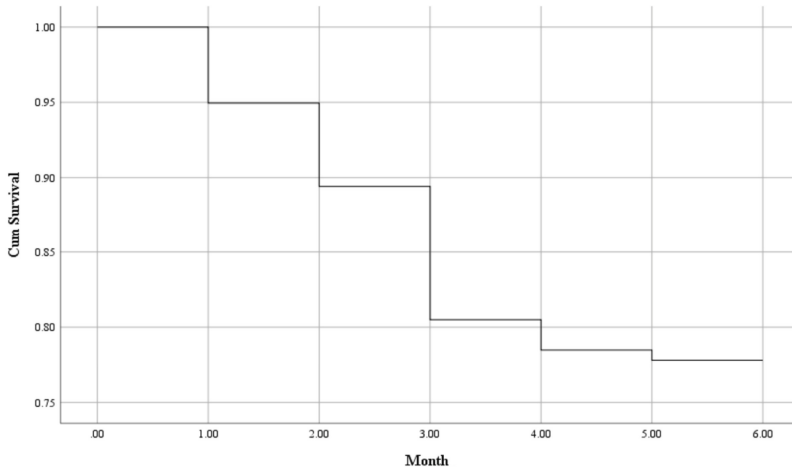
**Table 3.** Survival status of 420 service firms during COVID-19

	Number of companies operating by the end of 2020 (a)	Number of companies probably fail before six months (b)	Failure rate (b/a)	Nelson-Aalen cumulative hazard function			
				Coef	S.E	(95% confidence interval)	
All companies in the service sector	420	106	0.252	0.279	0.028	0.238	0.348
Business breakdowns (TSIC)							
Construction	118	12	0.102	0.105	0.031	0.060	0.188
Travel agency	28	10	0.357	0.411	0.136	0.228	0.796
Personal care	19	2	0.105	0.108	0.076	0.027	0.433
Logistics	20	3	0.150	0.155	0.091	0.051	0.491
Passenger land transport	9	1	0.111	0.111	0.111	0.015	0.789
Entertainment	12	4	0.333	0.367	0.194	0.144	1.033
Accommodation	56	31	0.554	0.725	0.144	0.544	1.121
Restaurant	51	13	0.255	0.284	0.080	0.167	0.497
Real estate	107	30	0.284	0.312	0.059	0.226	0.463

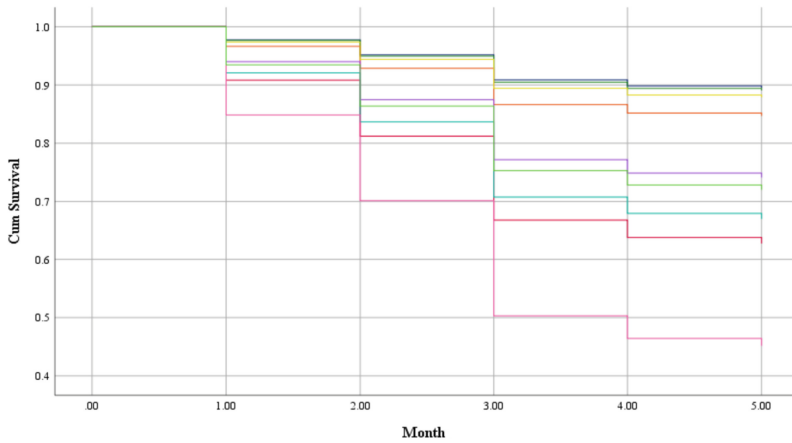
Figure 1 illustrates the differences in survival probability of the service sector depending on the categories. Among the nine businesses investigated, accommodation had the highest failure rate at 0.554, while construction had the lowest failure risk at 0.102. The risk of the firm’s failure in the other seven sub-sectors ranged from 0.105 to 0.357. The long-rank test for the survival curve of the business categories rejected the null hypothesis ( $\chi^2 = 55.32$ ,  $MBF \leq 0.001$ ), demonstrating that different service sectors were affected by COVID-19 differently. The Minimum Bayes Factor (MBF) value provides the strength of evidence against the null hypothesis (the difference in the mean is equal to zero). It is the smallest possible Bayes Factor for the point null hypothesis against the alternative within the specified class of alternatives [42]. In light of this, the semi-parametric Cox proportional hazard model was used to estimate the outcomes.

Table 4 depicts the results of the three versions of the semi-parametric Cox proportional hazard model. For model 1, only the firm characteristic variables were included in the regression. Model 2 covered the firm characteristic variables, financial variables, and government support, except for the business categories. Model 3 considered all independent variables to test the relationship between the dependent and independent variables. The following hypotheses’ results were mainly related to Model 3.

This study hypothesized that firm size and age had a positively favorable effect on firm survival. However, our results revealed that firm size and age are not found to affect business survival probability in both linear and non-linear effects. This finding may appear unexpected at first because firm size and age are usually positively associated with firm survival. As a result of the study, we



(a) All firms



(b) Service sector breakdown by business categories (TSIC)

**Fig. 1.** Kaplan-Meier survival function of the business categories **Notes:** This graph depicts the time series shifts in the survival function of the business categories based on the Kaplan-Meier estimator’s calculation. The vertical axis represents the likelihood of business survival, whereas the horizontal axis represents the observation duration. In Panel (b), the service sectors represent the nine business categories: construction (blue), personal care (dark green), passenger land transport (yellow), logistics (orange), restaurant (purple), real estate (green), entertainment (sky blue), travel agency (red), and accommodation (pink). The long-rank test for equality of survival function: ( $\chi^2 = 55.32$ ,  $MBF \leq 0.001$  at  $p \leq 0.0001$ ) [33].

believe that the COVID-19 pandemic was a novel exogenous shock which meant that most SMEs were unable to predict and unprepared for the shock [47]. In addition, Baumöhl et al. [11] concluded that there was some ambiguity about the impacts of firm size and age. Hence, the hypothesis that “firm size and age affect firm survival” was rejected.

The legal structure of a corporation classified as a partnership negatively influenced the chance of survival since the hazard ratio was significantly higher than 1. This finding was consistent with a study by Harhoff et al. [32], which reported that limited liability companies were more likely to collapse than unlimited liability companies. However, the legal structure may have a distinct influence on the chance of survival; for example, Esteve-Pérez and Maez-Castillejo [24] discovered that limited liability companies outlived other types of companies. As a result, the hypothesis that “the legal structure of a corporation influences firm survival” is correct.

The risk of company failure varied according to the business categories. Our findings indicated that four types of businesses (i.e., travel agency, entertainment, accommodation, and restaurant) had the highest chance of failure compared to the baseline categories. The hazard ratio exceeding one can be explained by the fact that the business is mostly concerned with tourism-related activities. Thus, the probability of firm survival in the service sector depends on the intensity of tourism activities [53]. Hence, we do not reject our hypothesis that the business category affects the firm survival.

The financial statement demonstrated economically significant factors affecting firm survival because the hazard ratio was either significantly higher or much lower than the threshold of 1. The preventive factor on firm survival was based on statistical significance. The results showed that three covariates had a significant impact on firm survival. Liquidity was a substantial factor in increasing firm survival. These results were in line with a study by Brown et al. [14] and Holtz-Eakin et al. [35], which revealed that lack of liquidity led a vulnerable firm to external (economic) shocks. Thus, cash flow management is an important strategy for survival during a critical situation. The coefficient, however, had a neutral effect on firm survival because it oscillated around the threshold of 1.

The operation efficiency was found to be a statistically significant factor affecting firm survival because the hazard ratio was below 1. The results showed that a firm with poor efficiency had a high probability of failure, although the profitability (ROA) was insignificant. The ROA finding may seem unexpected initially because profit was frequently linked to a firm’s ability to survive. Thus, this might hint that, in driving firm survival in an economic crisis, the ability to generate reasonable revenue is more important than profit. The findings in this work agreed with previous empirical research, which suggested that a firm’s operating income was a key determinant for firm survival [21].

**Table 4.** The Cox proportional hazard model's estimation results

Model Variable	[1]		[2]		[3]	
	Haz. Ratio	MBF	Haz. Ratio	MBF	Haz. Ratio	MBF
Size	1.211 (0.144)	0.2038	1.091 (0.137)	0.6940	0.986 (0.152)	0.9922
Age	1.006 (0.184)	0.9991	1.079 (0.259)	0.9251	1.118 (0.233)	0.8139
lnSize	0.464 (0.618)	0.2497	0.882 (0.613)	0.9630	1.203 (0.648)	0.9293
lnAge	1.362 (1.128)	0.9347	0.819 (1.423)	0.9824	0.722 (1.327)	0.9472
Size <sup>2</sup>	0.995 (0.003)	0.0680	0.997 (0.003)	0.3006	0.998 (0.003)	0.6512
Age <sup>2</sup>	0.999 (0.003)	0.8984	0.998 (0.005)	0.8433	0.997 (0.004)	0.6018
Partnership	1.242 (0.286)	0.5958	1.566 (0.269)	0.0822	2.106 (0.271)	*** 0.0011
Limited liability company	1.366 (0.231)	0.1936	1.223 (0.232)	0.5069	1.335 (0.244)	0.2825
Travel agency	4.731 (0.434)	*** 0.0000			6.070 (0.369)	*** 0.0000
Personal care	1.085 (0.774)	0.9900			1.368 (0.834)	0.8806
Logistics	1.447 (0.647)	0.7456			0.663 (0.725)	0.7490
Passenger land transport	1.081 (1.105)	0.9955			0.720 (1.161)	0.9305
Entertainment	3.590 (0.591)	* 0.0149			3.126 (0.577)	* 0.0298
Accommodation	7.493 (0.354)	*** 0.0000			2.902 (0.393)	*** 0.0014
Restaurant	2.894 (0.401)	*** 0.0018			2.684 (0.372)	*** 0.0017
Real estate	2.816 (0.346)	*** 0.0003			1.300 (0.355)	0.6110
Profitability			0.996 (0.002)	* 0.0241	0.998 (0.002)	0.3285
Liquidity			0.970 (0.013)	** 0.0056	0.966 (0.014)	** 0.0048
Efficiency			0.530 (0.248)	*** 0.0027	0.611 (0.209)	** 0.0068
Leverage			1.134 (0.074)	0.0759	1.191 (0.059)	*** 0.0004
Government support			0.460 (0.195)	*** 0.0000	0.470 (0.226)	*** 0.0000
<i>N</i>	420		420		420	
Log pseudolikelihood	-598.07		-550.78		-537.13	
Harrell's C-statistic	0.716		0.8286		0.8447	
Wald test ( $\chi^2$ )	56.98	***	95.21	***	135.17	***

**Notes:** The standard errors are calculated using the Huber-White sandwich estimator. The Minimum Bayes Factor (MBF) provides the strength of evidence against the null hypothesis. \*\*\*, \*\*, and \* represent decisive, very strong, and strong evidence, respectively.

As the predicted hazard ratio exceeded the threshold of 1, the degree of debt had a substantial link with firm survival, indicating that high levels of leverage enhanced the chance of a firm failure. This result was consistent with a study by Cresp-Cladera et al. [21], which reported that firms with greater leverage were more likely to fail during an economic recession. As a result, the hypothesis that “leverage influences survival probability” is correct.

Finally, a firm might survive through the economic crisis via government support. The degree of government support is then determined by policy inclusion. Our results indicated that the government support was economically significantly preventing firm failure. The magnitude of the hazard ratio was high as the coefficient was less than the threshold of 1. This finding aligned with the fact that SMEs found it extremely challenging to withstand the COVID-19 crisis without the government’s support [48]. As a robustness check, this paper performed the alternative parametric model by re-estimating the Cox proportional hazard model with various distributional hypotheses for survival functions, comprising the Exponential, Weibull, Gompertz, Log-normal, and Log-logistic (see Table 5).

## 5 Conclusion

This study examines the survival probability of firms in the service sector during COVID-19 by employing firm-level data from nine business categories (travel agency, personal care, logistics, passenger land transportation, entertainment, accommodation, restaurant, and real estate). Based on the Cox proportional hazard model, this work empirically discovered that different business categories faced distinct failure risks and experimentally identified the factors impacting the likelihood of firm survival.

The study surveyed 420 small- and medium-sized enterprises in Chiang Mai, Thailand, in 2021. Of all enterprises surveyed, 106 enterprises probably faced failure six months after the COVID-19 pandemic. The results revealed that the pandemic caused an economic shock throughout the service sector. The Cox proportional hazard model results showed that the severity of the economic crisis differed substantially between firms, notwithstanding the complexity of their dimensions. Firm activity directly related to tourism was highly vulnerable to pandemics, depending on the intensity of business tourism activities [53].

Furthermore, the empirical findings of the survival study showed that liquidity and efficiency had a beneficial impact on the chance of survival. This result was consistent with previous studies on this sector. However, the findings demonstrated that leverage had a detrimental influence on firm longevity, meaning that a firm with a larger degree of debt may be unable to produce income and face collapse. Thus, the government and banks played a significant role in increasing firm survival during the pandemic by providing loans to offset short-term liquidity concerns [56].

The specific policy support highlighted by the sample SMEs is financial support in the form of soft loans, suspension of interest rates, deferral of payment, and rollover of debt, in addition to tax preferences such as tax reduction and deferrals. Particularly, service firms relevant to tourism (travel agencies, entertainment, accommodation, and restaurant) requested operational subsidies, including employment cost support, office rent, and abatement of daily expenses for water and electricity. In the workplace, there was a substantial demand for supplies for pandemic protection, including masks and disinfectants. Most SMEs favored social insurance exemptions for employment subsidies; while for government services, the SMEs largely demanded specialized service improvements connected to the epidemic. As a result, it became clear that the government would need to develop tailored policies based on the unique characteristics of the various businesses affected by the pandemic. Given the impact on unemployment, policymakers tend to focus on preventing service industries that are heavily invested in tourism. However, the period of epidemic prevention measures has significantly impacted the firm's survival. Thus, removing pandemic-related legislation limits leads to the economic recovery of some businesses. This research gives guidance for developing a strategy to assist enterprises in dealing with the pandemic shock. If the pandemic reoccurs, we hope the impacts will be less severe and easier to contain if the lessons of the recent epidemic are followed.



# A Appendix

**Table 5.** Estimation results of parametric survival function

Model	[1]	[2]	[3]	[4]	[5]	[6]
Assumption of survival function	Cox proportional hazard	Exponential	Weibull	Gompertz	Log-normal	Log logistic
Size	0.986 (0.152)	1.013 (0.162)	1.013 (0.155)	1.014 (0.163)	-0.047 (0.122)	-0.039 (0.131)
Age	1.118 (0.233)	1.131 (0.248)	1.140 (0.268)	1.125 (0.237)	-0.061 (0.174)	-0.077 (0.236)
lnSize	1.203 (0.648)	1.043 (0.667)	1.028 (0.700)	1.060 (0.646)	0.056 (0.545)	0.036 (0.571)
lnAge	0.722 (1.327)	0.685 (1.394)	0.650 (1.493)	0.708 (1.337)	-0.109 (1.022)	0.007 (1.276)
Size <sup>2</sup>	0.998 (0.003)	0.998 (0.003)	0.998 (0.003)	0.998 (0.003)	0.002 (0.002)	0.002 (0.003)
Age <sup>2</sup>	0.997 (0.004)	0.997 (0.004)	0.997 (0.005)	0.997 (0.004)	0.001 (0.003)	0.002 (0.005)
Partnership	2.106*** (0.271)	1.979** (0.270)	2.100*** (0.290)	1.920** (0.263)	-0.512** (0.212)	-0.527** (0.227)
Limited liability company	1.335 (0.244)	1.350 (0.253)	1.356 (0.264)	1.345 (0.245)	-0.367 (0.202)	-0.373 (0.221)
Travel agency	6.070*** (0.369)	6.200*** (0.386)	6.729*** (0.406)	5.884*** (0.376)	-1.559*** (0.344)	-1.506*** (0.338)
Personal care	1.368 (0.834)	1.280 (0.849)	1.269 (0.863)	1.290 (0.839)	-0.467 (0.597)	-0.290 (0.756)
Logistic	0.663 (0.725)	0.653 (0.738)	0.620 (0.759)	0.674 (0.730)	0.051 (0.529)	0.185 (0.629)
Passenger land transport	0.720 (1.161)	0.681 (1.150)	0.658 (1.169)	0.696 (1.144)	-0.194 (0.938)	0.242 (0.999)
Entertainment	3.126* (0.577)	3.090* (0.621)	3.188* (0.649)	3.024* (0.601)	-1.007* (0.510)	-0.932* (0.543)
Accommodation	2.902*** (0.393)	2.727** (0.405)	2.849** (0.423)	2.662** (0.396)	-0.85*** (0.292)	-0.842*** (0.343)
Restaurant	2.684*** (0.372)	2.801*** (0.392)	2.914*** (0.402)	2.728*** (0.383)	-0.856*** (0.317)	-0.757** (0.359)
Real estate	1.300 (0.355)	1.265 (0.371)	1.234 (0.385)	1.282 (0.364)	-0.391 (0.265)	-0.335 (0.301)
Profitability	0.998 (0.002)	0.997 (0.002)	0.997 (0.002)	0.997 (0.002)	0.005 (0.002)	0.006 (0.004)
Liquidity	0.966** (0.014)	0.965** (0.015)	0.963** (0.002)	0.965** (0.009)	0.022*** (0.006)	0.025*** (0.009)
Efficiency	0.611** (0.209)	0.596** (0.225)	0.577** (0.016)	0.607** (0.215)	0.410*** (0.128)	0.398*** (0.156)
Leverage	1.191*** (0.059)	1.183*** (0.057)	1.203*** (0.235)	1.171*** (0.054)	-0.134** (0.055)	-0.134*** (0.045)
Government support	0.470*** (0.226)	0.459*** (0.229)	0.453*** (0.062)	0.462*** (0.225)	0.758*** (0.170)	0.744*** (0.210)
Constant	-	0.030**	0.020***	0.034**	3.70***	3.48***
N	420	420	420	420	420	420
Log pseudolikelihood	-537.13	-262.71	-261.15	-262.12	-253.91	-255.79
Wald test ( $\chi^2$ )	135.17***	144.5***	124.3***	139.3***	169.3***	170.2***

**Notes:** Results from the survival function using five parametric estimators are shown in this table as a robustness check. The Models [1] through [4] provide hazard ratios, whereas Models [5] through [6] provide regression coefficients. The Wald test examines the null hypothesis that all coefficients are zero. The Minimum Bayes Factor (MBF) provides the strength of evidence against the null hypothesis. \*\*\*, \*\*, and \* denote decisive, very strong, and strong evidence, respectively.

## References

1. Acs, Z.J., Audretsch, D.B.: Births and firm size. *Southern Econ. J.* 467–475 (1989)
2. Adam, N.A., Alarifi, G.: Innovation practices for survival of small and medium enterprises (SMEs) in the COVID-19 times: the role of external support. *J. Innov. Entrep.* **10**(1), 1–22 (2021). <https://doi.org/10.1186/s13731-021-00156-6>
3. Agarwal, R.: Small firm survival and technological activity. *Small Bus. Econ.* **11**(3), 215–224 (1998)
4. Agarwal, R., Audretsch, D.B.: Does entry size matter? The impact of the life cycle and technology on firm survival. *J. Ind. Econ.* **49**(1), 21–43 (2001)
5. Agarwal, R., Gort, M.: The evolution of markets and entry, exit and survival of firms. *Rev. Econ. Stat.* 489–498 (1996)
6. Audretsch, D.B.: New-firm survival and the technological regime. *Rev. Econ. Stat.* 441–450 (1991)
7. Audretsch, D.B., Mahmood, T.: The rate of hazard confronting new firms and plants in US manufacturing. *Rev. Ind. Organ.* **9**(1), 41–56 (1994)
8. Audretsch, D.B., Santarelli, E., Vivarelli, M.: Start-up size and industrial dynamics: some evidence from Italian manufacturing. *Int. J. Ind. Organ.* **17**(7), 965–983 (1999)
9. Azad, N.F., Serletis, A., Xu, L.: COVID-19 and monetary-fiscal policy interactions in Canada. *Q. Rev. Econ. Finance* **81**, 376–384 (2021)
10. Bartik, A.W., Bertrand, M., Cullen, Z.B., Glaeser, E.L., Luca, M., Stanton, C.T.: How are small businesses adjusting to COVID-19? Early evidence from a survey (No. w26989). National Bureau of Economic Research (2020)
11. Baum, J.A.: Liabilities of newness, adolescence, and obsolescence: exploring age dependence in the dissolution of organizational relationships and organizations. *Proc. Adm. Sci. Assoc. Canada* **10**(5), 1–10 (1989)
12. Borio, C.: The COVID-19 economic crisis: dangerously unique. *Bus. Econ.* **55**(4), 181–190 (2020)
13. Bougheas, S., Mizen, P., Yalcin, C.: Access to external finance: theory and evidence on the impact of monetary policy and firm-specific characteristics. *J. Bank. Financ.* **30**(1), 199–227 (2006)
14. Brown, R., Rocha, A., Cowling, M.: Financing entrepreneurship in times of crisis: exploring the impact of COVID-19 on the market for entrepreneurial finance in the United Kingdom. *Int. Small Bus. J.* **38**(5), 380–390 (2020)
15. Bruderl, J., Schussler, R.: Organizational mortality: the liabilities of newness and adolescence. *Adm. Sci. Q.* 530–547 (1990)
16. Çakmak, E., Lie, R., McCabe, S.: Reframing informal tourism entrepreneurial practices: capital and field relations structuring the informal tourism economy of Chiang Mai. *Ann. Tour. Res.* **72**, 37–47 (2018)
17. Carroll, G., Hannan, M.T.: *The Demography of Corporations and Industries*. Princeton University Press, Princeton (2000)
18. Chen, K.C., Lee, C.W.: Financial ratios and corporate endurance: a case of the oil and gas industry. *Contemp. Account. Res.* **9**(2), 667–694 (1993)
19. Clementi, G.L., Hopenhayn, H.A.: A theory of financing constraints and firm dynamics. *Q. J. Econ.* **121**(1), 229–265 (2006)
20. Cox, D.R.: Regression models and life-tables. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **34**(2), 187–202 (1972)
21. Crespi-Cladera, R., Martín-Oliver, A., Pascual-Fuster, B.: Financial distress in the hospitality industry during the COVID-19 disaster. *Tour. Manage.* **85**, 104301 (2021)

22. De Vito, A., Gómez, J.P.: Estimating the COVID-19 cash crunch: global evidence and policy. *J. Account. Public Policy* **39**(2), 106741 (2020)
23. Ericson, R., Pakes, A.: Markov-perfect industry dynamics: a framework for empirical work. *Rev. Econ. Stud.* **62**(1), 53–82 (1995)
24. Esteve-Pérez, S., Mañez-Castillejo, J.A.: The resource-based theory of the firm and firm survival. *Small Bus. Econ.* **30**(3), 231–249 (2008)
25. Eurostat: Tourism Industries - Economic Analysis. Eurostate - Statistics Explained (2018)
26. Fichman, M., Levinthal, D.A.: Honeymoons and the liability of adolescence: a new perspective on duration dependence in social and organizational relationships. *Acad. Manag. Rev.* **16**(2), 442–468 (1991)
27. Fotopoulos, G., Louri, H.: Location and survival of new entry. *Small Bus. Econ.* **14**(4), 311–321 (2000)
28. Freeman, J., Carroll, G.R., Hannan, M.T.: The liability of newness: age dependence in organizational death rates. *Am. Sociol. Rev.* 692–710 (1983)
29. Geroski, P.A., Mata, J., Portugal, P.: Founding conditions and the survival of new firms. *Strateg. Manag. J.* **31**(5), 510–529 (2010)
30. Görg, H., Spaliara, M.E.: Financial health, exports and firm survival: evidence from UK and French firms. *Economica* **81**(323), 419–444 (2014)
31. Hannan, M.T.: Rethinking age dependence in organizational mortality: logical formalizations. *Am. J. Sociol.* **104**(1), 126–164 (1998)
32. Harhoff, D., Stahl, K., Woywode, M.: Legal form, growth and exit of West German firms-empirical results for manufacturing, construction, trade and service industries. *J. Ind. Econ.* **46**(4), 453–488 (1998)
33. Held, L., Ott, M.: On p-values and Bayes factors. *Ann. Rev. Stat. Appl.* **5**(1), 393–419 (2018)
34. Hitt, M.A., Ireland, R.D., Hoskisson, R.E.: *Strategic Management: Concepts and Cases: Competitiveness and Globalization*. Cengage Learning (2016)
35. Holtz-Eakin, D., Joulfaian, D., Rosen, H.S.: Sticking it out: entrepreneurial survival and liquidity constraints. *J. Polit. Econ.* **102**(1), 53–75 (1994)
36. Jovanovic, B.: Selection and the Evolution of Industry. *Economet.: J. Economet. Soc.* 649–670 (1982)
37. Kumar, P.R., Ravi, V.: Bankruptcy prediction in banks and firms via statistical and intelligent techniques-a review. *Eur. J. Oper. Res.* **180**(1), 1–28 (2007)
38. Lane, W.R., Looney, S.W., Wansley, J.W.: An application of the Cox proportional hazards model to bank failure. *J. Bank. Financ.* **10**(4), 511–531 (1986)
39. Levinthal, D.A.: Adaptation on rugged landscapes. *Manage. Sci.* **43**(7), 934–950 (1997)
40. Lu, L., Peng, J., Wu, J., Lu, Y.: Perceived impact of the COVID-19 crisis on SMEs in different industry sectors: evidence from Sichuan, China. *Int. J. Disast. Risk Reduct.* **55**, 102085 (2021)
41. Manjón-Antolín, M.C., Arauzo-Carod, J.M.: Firm survival: methods and evidence. *Empirica* **35**(1), 1–24 (2008)
42. Maneejuk, P., Yamaka, W.: Significance test for linear regression: how to test without P-values? *J. Appl. Stat.* **48**(5), 827–845 (2021)
43. Martinez, M.G., Zouaghi, F., Marco, T.G., Robinson, C.: What drives business failure? Exploring the role of internal and external knowledge capabilities during the global financial crisis. *J. Bus. Res.* **98**, 441–449 (2019)
44. Mata, J., Portugal, P.: Life duration of new firms. *J. Industr. Econ.* 227–245 (1994)
45. Mata, J., Portugal, P., Guimaraes, P.: The survival of new plants: start-up conditions and post-entry evolution. *Int. J. Ind. Organ.* **13**(4), 459–481 (1995)

46. McGeever, N., McQuinn, J.: SME liquidity needs during the COVID-19 shock (No. 2/FS/20). Central Bank of Ireland (2020)
47. Miklian, J., Hoelscher, K.: SMEs and exogenous shocks: a conceptual literature review and forward research agenda. *Int. Small Bus. J.* **40**(2), 178–204 (2022)
48. Najib, M., Abdul Rahman, A.A., Fahma, F.: Business survival of small and medium-sized restaurants through a crisis: the role of government support and innovation. *Sustainability* **13**(19), 10535 (2021)
49. Omar, A.R., Ishak, S., Jusoh, M.A.: The impact of COVID-19 movement control order on SMEs' businesses and survival strategies. *Geografia* **16**(2) (2020)
50. Page, S., Yeoman, I., Munro, C., Connell, J., Walker, L.: A case study of best practice—Visit Scotland's prepared response to an influenza pandemic. *Tour. Manage.* **27**(3), 361–393 (2006)
51. Pu, G., Qamruzzaman, M., Mehta, A.M., Naqvi, F.N., Karim, S.: Innovative finance, technological adaptation and SMEs sustainability: the mediating role of government support during COVID-19 pandemic. *Sustadinability* **13**(16), 9218 (2021)
52. Puttachai, W., Yamaka, W., Maneejuk, P., Sriboonchitta, S.: Analysis of the global economic crisis using the cox proportional hazards model. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *ECONVN 2019. SCI*, vol. 809, pp. 863–872. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_62](https://doi.org/10.1007/978-3-030-04200-4_62)
53. Reis, H., Rodrigues, P.M., Caires, F.B.: Survival of the fittest: tourism exposure and firm survival (No. w202206) (2022)
54. Rico, M., Pandit, N.R., Puig, F.: SME insolvency, bankruptcy, and survival: an examination of retrenchment strategies. *Small Bus. Econ.* **57**(1), 111–126 (2021)
55. Robinson, J., Kengatharan, N.: Exploring the effect of COVID-19 on small and medium enterprises: early evidence from Sri Lanka. *J. Appl. Econ. Bus. Res.* **10**(2), 115–124 (2020)
56. Şengel, Ü., Işkın, M., Çevrimkaya, M., Genç, G.: Fiscal and monetary policies supporting the tourism industry during COVID-19. *J. Hospit. Tour. Insights* (2022)
57. Siriphattasophon, S.: The COVID-19 pandemic: impacts on Thai small and mediums enterprises and strategies for revival. In: *Impacts and Strategies for Tourism, Service Management, and Academic Sector, India* (2020)
58. Tsoukas, S.: Firm survival and financial development: evidence from a panel of emerging Asian economies. *J. Bank. Financ.* **35**(7), 1736–1752 (2011)
59. Wilder-Smith, A.: The severe acute respiratory syndrome: impact on travel and tourism. *Travel Med. Infect. Dis.* **4**(2), 53–60 (2006)



# How Non – Interest Income Matters for Operation Efficiency? A Bayesian Analysis of Vietnam Banks

Bui Dan Thanh<sup>1</sup>(✉), Doan Thanh Ha<sup>1</sup>(✉), and Pham Thi Hong Nhung<sup>2</sup>(✉)

<sup>1</sup> Banking University HCMC, Ho Chi Minh City, Vietnam  
{ thanhbd, hadt }@buh.edu.vn

<sup>2</sup> Ho Chi Minh City College of Economics, Ho Chi Minh City, Vietnam  
hongnhunghce2911@gmail.com

**Abstract.** This study tries to determine how non-interest revenue may affect 30 commercial banks in Vietnam's probability ratio between 2011 and 2020. Using the Bayesian regression technique, secondary data from 30 commercial banks was used to determine the effect of non-interest income on the probability ratio. Regression analysis reveals that non – interest revenue, bank size, debt to equity ratio, operational expenses, deposit rate, and inflation have a positive and statistically significant impact on the operating performance of Vietnamese commercial banks. In contrast, neither the GDP growth rate nor the provision for bad debts have a statistically significant impact on the profitability of Vietnam's commercial banks.

**Keywords:** commercial banks · non-interest income · operational efficiency

## 1 Introduction

In recent years, the world banking industry has had great changes in technology, competitive environment, customer needs, etc. These are the reasons for the continuous development of products and non-traditional service. The expansion of policies as well as forms of non-traditional financial products and services, combined with attractive interest rates to attract customers, are being used by banks to become more popular, richer and more attractive. The bank's non-interest income comes from service fees, commissions, insurance, securities, etc. From here, banks can attract a large number of customers and increase their competitiveness against competitors in the same industry, increase the income of banks; and this is also an inevitable trend for the survival and development of commercial banks in developed countries.

Currently, in the world and in Vietnam, there are two sources of conflicting research results on the impact of non-interest income on the probability ratio of banks. According to research by Baele et al. (2007), thanks to the expansion of non-traditional activities, banks can have more opportunities to access new information sources, facilitate cross-selling of products, develop other products and services that have a more positive impact

on the business relationship. Some other studies also show a positive impact from non-interest income, non-traditional activities that help banks reduce risk such as Saunders, Schmid and Walter (2014), Singh and Upadhyay (2016), etc. Domestic studies also reached the same conclusion as Le Long Hau and Pham Xuan Quynh (2016), Ho Thi Hong Minh and Nguyen Thi Canh (2015), etc. However, there are also studies that give conflicting conclusions, when non-interest income increases risk as studied by Lepetit, Nys, Rous, and Tarazi (2008); Li and Zhang (2013); Williams (2016). They believe that the expansion of non-traditional business activities will lead to banks consuming large amounts of fixed costs, increasing operating leverage and higher risks. Because there are many multi-dimensional studies, from here, the question arises whether non-interest income really has a positive impact on the bank's probability ratio or vice versa, it will have an impact on the profitability of banks, and at the same time increase the risk of bankruptcy according to previous research by some authors. To answer and contribute to meet the practical ability to Vietnam, the author chooses to research the topic: "The impact of non-interest income on the probability ratio of Vietnamese commercial banks" to be able to provide a complete view and appropriate direction for commercial banks.

## 2 Literature Review

### 2.1 Theoretical Basis

#### 2.1.1 Operational Efficiency of Commercial Banks

According to Farrell (1957), efficiency represents the correlation between the obtained output variables and the input variables that were used to produce those outputs. In which, cost efficiency or economic efficiency includes technical efficiency and allocative efficiency. Technical efficiency is an indicator that represents the maximum production capacity of a unit with a given input, while allocative efficiency is used to reflect the ability to optimize inputs when prices are known. Therefore, efficiency can also be said to be the benefits brought from specific activities in the form of input optimization and output maximization.

According to the point of view expressed in the research papers of Elyasiani and Mehdian (1990a, 1990b) and Mester (1987), the output in the operation of commercial banks as financial intermediaries is the assets of banks, while the deposits, labor and capital are inputs. The most important thing in the bank's profit structure is interest income, and depends on the credit extension (namely lending). Therefore, credit development is very important for banks. In which, loan capital is considered as a product and loan interest rate is considered as the price of that credit.

According to Antonio, Ludger and Vito (2006), efficiency is seen as a comparison between inputs and outputs or between profits and costs. With the same level of input defined, the activity that produces more output is the more efficient operation. Efficiency is considered to be the degree to which firms or banks complete the allocation of usable inputs and the outputs they produce, meeting set goals (Nguyen Khac Minh 2004).

It can be understood, the operating system of a commercial bank is like the ability to generate profits of that commercial bank while limiting risks and all activities of the bank are still in accordance with the set goals. According to Chang et al. (2010) also

stated that efficiency also reflects the ability to manage, control costs and use resources to create outputs.

Commercial banks are an intermediary financial institution that plays a very important role in the market economy, there are always activities to transfer capital from places of excess to places of shortage (Peter 2014), but if viewed from the perspective of an enterprise, business activities must always have a combination of two factors, which is to maximize profits, but within the allowable risk level – the risk that commercial banks can accept.

Based on the above concepts, it can be concluded that the operating system of commercial banks is a collection of many factors such as financial resources, human resources, facilities and other factors in the operation of the bank. In the current integration context, focusing on the current business relationship is imperative and indispensable for each bank, contributing to creating the intrinsic value of each bank in the ever-changing market economy.

### 2.1.2 Income of Commercial Banks

**Net Interest Income:** Commercial banks provide capital mobilization products, loan products and a number of other business activities (domestic and international payments; domestic and foreign money transfer; salary collection and payment; SMS Banking; Internet Banking; trading in foreign currencies, gold and silver, cards; investing in government bonds, contributing capital to joint ventures; discounting commercial papers, bonds, valuable papers etc.). In which, lending activities play an important role in generating income of banks. The source of income from lending activities is called interest income – always accounts for a high proportion in all banks. Net interest income is the difference between the income from lending activities and the expenditure on capital mobilization activities. In other words, the gap between lending rates and deposit rates is net interest income (Hoang Ngoc Tien and Vo Thi Hien 2010). Net interest income depends on the interest rate management policy of each bank. Besides, other factors such as economic conditions, the flourishing of business lines, etc. are also factors affecting interest income. Interest income depends on the loan interest rate, outstanding balance and ability to cooperate in debt repayment. Although this source of income plays a key role in banking activities, it has many potential risks, because the loans depend on the economic situation of the customer, leading to the degree of cooperation in repayment of the customer. In addition, other factors such as economic conditions, the development of business lines, etc. are also factors affecting interest income.

**Non-Interest Income:** The calculation of income from non-interest business activities of commercial banks is based on the income structure including interest income (mainly coming from the bank's lending activities, this is the main source of income for the commercial banks) accounts for the largest proportion of total revenue of the bank and non-interest income. The source of income from activities other than credit activities is called non-interest income. According to Hoang Ngoc Tien and Vo Thi Hien (2010), private equity is being interested by banks in order to spread risks and reduce dependence on main sources of income, in which non-interest income is income from services, trading in foreign exchange, gold, silver, gems; securities trading and other service

activities. Stiroh (2004) argues that non-interest income is a heterogeneous category that includes many different activities, divided into four main components: trust income, service fees, fees. Therefore, based on the annual reports of banks, non-interest income includes: income from service activities, fees, commissions, income from investment business activities (foreign exchange, gold, trading and trading securities, investment securities trading, income from capital contribution, share purchase), other non-interest income. In general, the higher the ratio of non-interest income, the greater the bank's sensitivity to changes in interest rates.

### **2.1.3 The Impact of Non-interest Income on the Operating Performance of Commercial Banks**

Business relations are always a matter of concern to economic organizations as well as banks. The Bank's effective business operations demonstrate the Bank's capacity to customers and partners, and at the same time attract investors, thereby ensuring the stability and development of the Bank. Therefore, improving operational efficiency is always one of the top concerns of banks. The evaluation of the operational performance of commercial banks is not only of great significance for banks in considering the overall use of resources, enhancing competitiveness, but also for the management agencies of State in supporting and creating conditions for banks to operate better.

When non-interest income generating activities such as business, investment, payment services, card services are developed, commercial banks will make optimal and effective use of technical facilities, as well as human resources of each bank. Therefore, management costs and operating expenses are reduced, increasing the maximum profit for the bank. Expanding non-interest activities also helps banks to disperse and reduce risks, especially credit risks. Theoretically, in all activities that bring profits to banks, the activities of income diversification, increasing the ratio of non-interest income are very preferred by banks because of service fees, business profits. Net and non-interest income are not completely correlated with net interest income. Therefore, increasing non-interest income ratio leads to more stable operating income and better adjusted financial risk (Odesanmi and Wolfe 2007).

DeYoung and Rice (2004a) study the relationship between non-interest income and financial performance in the commercial banking industry in the US from 1989 to 2001. The results show that the rate of non-interest income has a positive effect on ROE. Similarly, Chiorazzo et al. (2008) using data of banks in Italy during the period 1993–2003, the authors also concluded that increasing non-interest income will increase the operational efficiency of banks, adjusting for systemic risk of banks in Italy and this relationship will be stronger in larger banks. According to a study by Baele et al. (2007) using panel data of banks in Europe from 1989 to 2004 it was concluded that an increase in non-interest income will increase the operational efficiency of banks. Besides, Busch and Kick (2009) also studied the impact of fee-based income on the profitability of banks from 1995 to 2007 in Germany. The results show that high fee-based income can increase profitability, adjust bank risk. Or according to Elsas et al. (2010) studied the impact of income diversification on both bank performance and market value using panel data of nine countries from 1996 to 2008. They found that income diversification can improve



bank profitability and market value. In addition, Lozano–Vivas and Pasiouras (2010), Gamra and Plihon (2011), Meslier et al. (2014) also argue that when the financial market is increasingly integrated, banks must strengthen their competitiveness through through a business diversification strategy. Since non–interest income generating businesses such as underwriting, brokerage and other consulting services often have a weak/unclear relationship with traditional credit operations, diversifying income sources will be a lifeline for the bank’s profits when lending activities go wrong.

Studies in Vietnam also come to similar conclusions as Le Long Hau and Pham Xuan Quynh (2016), Ho Thi Hong Minh and Nguyen Thi Canh (2015), Vo Xuan Vinh and Tran Thi Phuong Mai (2015).

However, there are also studies that show different perspectives and results. For example, Stiroh (2004) using data from commercial banks in the US in the period 1970 to 2001, the author concludes that non–interest income has a positive impact on the risk of insolvency of commercial banks, that is, is income diversification leading to commercial bank’s inability to pay. Or according to Mercieca et al. (2007) study whether non–interest income improves the performance of small credit institutions in Europe or not. Using a sample of 755 small banks in the period 1997–2003, the authors found an inverse relationship between non–interest income and bank performance. Besides, Lepetit et al. (2008) studied the relationship between banking risk and product diversification in the changing structure of the European banking industry. Based on a series of European banks between 1996 and 2002, our research shows that banks that expand into non–interest income activities have higher risk and higher risk of insolvency compared to banks. Banks mainly provide loans.

## 2.2 Comprehensive Researches

### 2.2.1 Comprehensive Researches in Vietnam

**Ho Thi Hong Minh and Nguyen Thi Canh (2015)** have focused on examining the relationship between the private diversification of banks and factors affecting profitability of Vietnamese commercial banks. This study uses data collected from the financial statements of 22 Vietnamese commercial banks in the period 2007–2013, processed through the SGMM method. The results show that the ratio of income diversification (the contribution ratio of non–interest income), the ratio of outstanding loans to total assets, the ratio of operating expenses to income are negatively correlated with profitability; In which, the expansion of non–profit business activities to help Vietnamese commercial banks increase profitability and develop service activities in parallel with credit activities is an inevitable development trend of Vietnamese commercial banks in the context of economic situation is increasingly difficult and competition is fierce.

**Le Long Hau and Pham Xuan Quynh (2016)** pointed out that the implementation of self–diversification by expanding products and services will have a positive impact on the bank’s business performance; at the same time, it also shows that net non–interest income sources, loan balance, equity size, bank size, economic growth rate, and inflation have a positive impact, while, operating costs, the activities and money of customers have a negative impact on the business performance of commercial banks. The article, through the processing of data collected from financial statements and annual reports

of 26 Vietnamese commercial banks in the period 2006–2014 using FEM and REM models also shows that the increase in the ratio of external income Interest rates are beneficial for Vietnamese commercial banks in addition to racing with credit growth (risky activities and high probability of bad debt). Therefore, in the current period of integration and increasingly fierce competition, the expansion of non–interest income generating activities, especially service activities, is necessary to increase income and improve efficiency business for commercial banks.

**Nguyen Minh Sang and Nguyen Thi Thuy Trang (2018)** analyzed the impact of non–interest income on the risk and profitability of 26 Vietnamese commercial banks from 2008 to 2016 using a panel data analysis model. The study uses the ratio of non–interest income to net operating profit to measure non–interest income. Research results show that non–interest income has no impact on risk but has a positive effect on profitability of commercial banks (typically, ROA) during the research period. This is a good sign for banks that want to diversify their private equity, especially non–interest income from non–traditional activities to improve competitiveness, limit risks and increase profits.

**Nguyen Thi Diem Hien and Nguyen Hong Ha (2016)** study on factors affecting non–interest income and its impact on financial performance of 33 joint stock commercial banks in the period 2006–2013. The results show that non–interest income has a positive impact on financial performance and reduces the volatility of financial performance. At the same time, the study points out that bank–specific factors and market conditions affect non–interest income.

**Trinh Thi Thuy Hong, Nguyen Hoang Phong and Le Tien Thanh (2018)** have shown that expanding non–interest business activities has a positive impact on business activities of state–owned commercial banks or in other words, private diversification has strong impact on state–owned commercial banks through annual panel data collected from a group of 29 Vietnamese commercial banks provided by Bankscope during the period from 2006 to 2016 with about 287 observations through the FEM and FGLS.

### 2.2.2 Comprehensive Researches in the World

**Al–Tarawneh, Abu Khalaf and Al Assaf (2017)** study the impact of non–interest income on the operations of 13 banks in Jordan during the period from 2000–2015 using a model FEM. The results indicate that bank size, loans, capital adequacy and overhead are found to have a significant impact on the performance of banks. Specifically, general costs reduce the bank’s operating efficiency, while the level of capital adequacy, loans and bank size increase the bank’s operating efficiency. In addition, non–interest income increases the safety of equity and this in turn has a positive effect on profitability.

**Bailey–Tapper (2010)** studied panel data of Jamaican commercial banks from March 1999 to September 2010 on the determinants of non–interest income. Research results show that: ATM technology, personal lending and loan quality are among the main microeconomic factors promoting non–interest income performance in the commercial banking sector. Regarding the macroeconomic environment, interest rates and exchange rate fluctuations are the main factors that explain the effect of non–interest income. In this context, the effect of stronger non–interest income not only increases profitability but also increases business efficiency. In addition, large banks have lower income in investment leading to increased service costs from loans and may reflect loans more

positively because these institutions increase fee income. In addition, lower income on investment also contributes to the higher “other” service costs of larger banks and it may reflect greater competitive demand, especially in the low interest rate environment of the period latter part of the Jamaican debt exchange.

**Chiorazzo, Milani and Salvini (2008)** used annual data from 85 Italian banks for the period 1993–2003, regression by FEM model to investigate the relationship between income non–interest income and profitability. The results show that the diversification of the private sector, the expansion of non–profit business types will increase the profits of banks. Small banks can benefit from increased non–interest income, but only if they have very little non–interest income to start a business. The study of Baele et al. (2007) also agrees with the above results. Accordingly, Baele et al. (2007) show that private banking helps banks to obtain customer information more easily, contributing to promoting cross–selling of products and enhancing other services.

**Craigwell and Maxwell (2005)** studied the trend of non–interest income at commercial banks in Barbados from 1985 to 2001, as well as study the determinants of non–interest income and its impact it affects the financial performance of commercial banks. The results indicate that the ratio of non–interest income in Barbados has decreased during this period. Analysis of the literature and panel data regression models suggest that the results for Barbados may be due to the absence of some of the most important factors in non–interest income generation in developed countries, such as deregulation and technological change, especially to develop loan securitization and credit risk assessment. Empirical evidence supports the characteristics of banks and ATM technology as the most influential factors shaping non–interest income trends in the banking industry in Barbados. In addition, non–interest income has a positive effect on both bank profitability and income volatility, indicating that non–interest income has a positive impact on bank profitability and income. Factors related to ATM technology affect non–interest income.

**DeYoung and Rice (2004a)** pointed out that non–interest income currently accounts for more than 40% of operating income for commercial banks in the US. The study also presents some empirical links between non–interest income of banks, business strategy, market conditions, technological change and financial performance from 1989 to 2001. The results show diversification to help banks reduce risks in business activities. Commercial banks take advantage of advanced technology for non–cash transactions to collect more fees to increase non–interest income. The study argues that non–interest income should co–exist with interest income from intermediary activities (rather than replacing or completely replacing), as these activities remain core financial services functions of banks.

**Oniang'o (2015)** argues that non–interest income is considered as an additional source of income for commercial banks, which is essential to improve profitability of commercial banks in Kenya. The study sought to determine the effect of non–interest income on the profitability of commercial banks in Kenya. To achieve this goal, the author used a descriptive survey. The subjects of the study included all 43 commercial banks in Kenya. Data analysis was performed using a regression model. Research shows that non–interest income has a positive impact on profitability of commercial banks. The results show that there is a moderate correlation between non–interest income and

profitability of commercial banks. The study recommends that companies should offset the risks of doing business. Similarly, the study by Som Raj Nepali (2018) examines the impact of private equity on the risk–return trade–off principle at Nepalese commercial banks, using secondary data collected from 20 Nepalese commercial banks since 2009 as well shows that non–interest income, foreign ownership ratio and bank size are positively correlated with risk–adjusted return.

**Tarazi, Crouzille and Tacneng (2010)** examine the impact of private diversification on the performance of banks in an emerging economy. The study shows that, in contrast to studies on the Western economy, the shift to non–interest–bearing activities increases banks’ profits and especially risk–adjusted returns when banks are more involved in trading in government securities. Foreign banks benefit more from this change than domestic banks. In addition, the study also takes into account the institutional and legal environment that supports loans to SMEs and finds that more participation in non–interest activities only benefits low–risk banks for SMEs.

**Trujillo–Ponce (2013)** empirically analyzed the determinants of profitability of Spanish banks in the period 1999–2009. The results show that the high profitability of the bank in these years is related to the large loan ratio in total assets, the high proportion of customer’s deposits, and the low ratio of bad assets. In addition, the study did not find statistical significance in the variable measuring the effect of income diversification on the profitability of banks, showing that non–interest income activities do not affect the profitability of banks in the Spain.

It is noteworthy that the research mentioned above used frequency approaches or descriptive analyses with suitable large sample sizes. This study, however, used Bayesian logistic regression with informative priors to assess how non–interest revenue may affect 30 commercial banks in Vietnam’s probability ratio between 2011 and 2020. The following contributions from the research have been as anticipated: Firstly, commercial banks will operate more efficiently if they move in the direction of increasing the proportion of non–interest income–generating businesses. Secondly, our results allow a broader conclusion that, in contrast to frequentist approaches, Bayesian estimation employing thoughtful priors can produce meaningful results. This generalization is made possible by using Bayesian MCMC simulations in informative (thoughtful) prior settings.

### 3 Models and Method

#### 3.1 Data Collection Methods

The study uses panel data, which is based on numerical data obtained from different sources. The first data source is from the audited financial statements and annual reports of thirty (30) commercial banks in Vietnam in the period 2011–2020 from the websites of commercial banks. The second data source is taken from the websites of international organizations, including open data of the World Bank, the General Statistics Office of Vietnam to ensure the reliability of the research.

#### 3.2 Research Models

From the point of view of inheriting and continuing to develop previous studies, the thesis uses the research model of Chiorazzo et al. (2008) and edited by Le Long Hau and

Pham Xuan Quynh (2017) to suit conditions of Vietnam. In fact, there are many previous studies that have chosen ROA to study the operational performance of banks. Therefore, in this study, the author chooses ROA as dependent variables. Besides the typical factors of the bank are non–interest income, size of bank, scale of credit activities, capital structure, operating costs, deposit size and macro factors such as economic growth rate and inflation; according to the studies of Weersainghe and Perera (2013); Sufian and Habibullah (2009); Ahmad (2014); Nguyen Thanh Phong (2015) can show that the asset quality of banks (LLP) is one of the variables that have an impact on the bank's operational performance. Realizing that the asset quality ratio of banks is mentioned in many studies, the author decided to add the asset quality variable of banks to improve the accuracy of the model. Therefore, the proposed model for this study is as follows:

$$ROA_{i,t} = \alpha + \beta_1 * ICO_{NON_{i,t}} + \beta_2 * SIZE_{i,t} + \beta_3 * LOAN_{i,t} + \beta_4 * EQUITY_{i,t} + \beta_5 * COST_{i,t} + \beta_6 * DTL_{i,t} + \beta_7 * LLP_{i,t} + \beta_8 * GDP_t + \beta_9 * INF_t + u_{i,t}$$

In which:

**ICO<sub>NON</sub> – Non–Interest Income Ratio:** According to Chronopoulos et al. (2011), Elyasiani and Wang (2012), Abdul (2015), Chiorazzo et al. (2008), Stiroh and Rumble (2006), Ho Thi Hong Minh and Nguyen Thi Canh (2015), Le Long Hau and Pham Xuan Quynh (2017), the ratio of non–interest income of commercial banks is estimated by the formula:

$$\begin{aligned} ICO_{NON} &= (ICO_{COM} + ICO_{TRAD} + ICO_{OTH}) \times 100(\%) \\ &= \left( \frac{COM}{NET + NON} + \frac{TRAD}{NET + NON} + \frac{OTH}{NET + NON} \right) \times 100(\%) \\ &= \frac{COM + TRAD + OTH}{NET + NON} \times 100(\%) \\ &= \frac{COM + TRAD + OTH}{NETOP} \times 100(\%) \end{aligned}$$

In which:

- ICO<sub>NON</sub>: ratio of non–interest income
- ICO<sub>COM</sub>: percentage of net income from service activities
- ICO<sub>TRAD</sub>: ratio of net income from business activities, investment
- ICO<sub>OTH</sub>: ratio of net income from other non–interest activities
- COM: net income from service activities
- TRAD: net income from business activities, investment
- OTH: net income from other non–interest activities
- NON: net income other than interest
- NET: net interest income
- NETOP: total income of commercial banks (NETOP = NON + NET)

If net non-interest income is negative, it is considered as  $ICO_{NON} = 0\%$ , which means that the bank does not use private equity (Ho Thi Hong Minh and Nguyen Thi Canh 2015). Thereby, the higher the  $ICO_{NON}$  index, the higher the contribution ratio of non-interest income of banks. The study expects banks to promote and maximize the productivity of non-traditional business activities (ie, the higher the  $ICO_{NON}$ ), the higher the operational efficiency of commercial banks. **Hypothesis 1: Non-interest income (ICONON) has a positive impact on the operating performance of commercial banks ( $H_1$ ).**

**SIZE – Bank size:** Bank size is measured by taking the logarithm of total assets of that bank (Nguyen Minh Kieu 2009). The data are placed in logarithmic form because this is a strong trending feature and it dominates the rest of the components (Pham Thi Tuyet Trinh 2016). Large banks will generate revenues from related services (Elsas et al. 2010; Chiorazzo et al. 2008). **Hypothesis 2: The size of the bank (SIZE) has a positive impact on the operating performance of commercial banks ( $H_2$ ).**

**LOAN – Size of Credit Activity:** Measured by taking the ratio of outstanding loans to total assets (Mercieca et al. 2007). This is an indicator showing the impact of lending activities on the bank's business performance (Chiorazzo et al. 2008). Bank profits will increase as loans increase, meaning that banks are focusing on lending rather than paying attention to other activities. **Hypothesis 3: The size of credit activities (LOAN) has a positive impact on the operating performance of commercial banks ( $H_3$ ).**

**EQUITY – Capital Structure:** Measured by the ratio of equity to total assets (Abd Karim et al. 2010). Theoretically, a bank's capital ratio is often tied to its own size because large banks tend to generate more profits than small banks due to their less expensive ability to raise capital. Bourke (1989) and Nguyen Tran Thinh (2013) have shown that the higher the capital ratio, the more profitable banks will be. In addition, an increase in the capital ratio can also provide unexpected returns from anticipated cost reductions from economic risks (including bankruptcy) according to Berger's research (1995) and was retested according to the study of Sufian, F. (2011). **Hypothesis 4: Capital structure (EQUITY) has a positive impact on the performance of commercial banks ( $H_4$ ).**

**COST – Operating Expenses:** Measured by operating expenses on total assets Obamuyi (2013). This ratio shows the total cost of the bank as a percentage of the total assets. The higher this ratio shows that the bank has not managed its expenses effectively and vice versa, the low this ratio shows that the bank's expenses are well managed, showing the talent and vision of the bank's managers. The higher the cost, the lower the profit. On the contrary, when banks manage costs well, profits will increase significantly. From there, costs will have a negative impact on bank profits. According to Guru et al. (2002); Bourke (1989) argues that the bank that cuts costs and uses management costs effectively, the more efficient it is. **Hypothesis 5: Operating cost (COST) has a negative impact on the performance of commercial banks ( $H_5$ ).**

**DTL – Deposit Size:** Measured by customer deposits over total liabilities (Shiers 2002). In which the total liabilities of the bank, the capital from customer deposits is said to be a stable and cheaper source of funding compared to other sources of funding (Ho Thi Hong Minh and Nguyen Thi Canh 2015). Thus, the higher the customer deposit ratio,

the higher the bank's profitability. **Hypothesis 6: Deposit size (DTL) has a positive impact on the performance of commercial banks (H<sub>6</sub>).**

**LLP – Bank's Asset Quality:** Calculated according to the ratio of provision for bad debts to total loans to customers (Berger et al. 2010). This ratio reflects what percentage of the loan balance is provisioned. The higher this index, the higher the quality of the credits of that bank is not good, the debt collection ability is low or vice versa, when this ratio is low, the quality of the credits is improved positively or maybe the provisions are not set up in accordance with regulations. Meanwhile, credit is the most profitable activity for banks, so this factor is expected to negatively affect profitability. Weersainghe and Perera (2013) show that the asset quality of banks has a negative impact on the profitability of banks. **Hypothesis 7: Asset quality (LLP) has a negative impact on the performance of commercial banks (H<sub>7</sub>).**

**GDP – Economic Growth Rate:** The author uses GDP growth rate to control for macroeconomic cycles. The economic growth rate is measured by the nominal GDP growth rate:

$$\text{Economic growth rate} = \frac{\text{GDP}_t - \text{GDP}_{t-1}}{\text{GDP}_{t-1}} \times 100(\%)$$

Source: Delis, M. D. (2012)

In the context of the unstable economy, the Bank will tighten loans, limit lending and reduce deposit interest rates, in addition, increase the provision for credit risks and reduce the bank's profitability. On the contrary, if the economic situation is positive, people's incomes will be stable, so the amount of savings flowing into banks will increase and businesses expand their investment with the need to borrow capital and thereby increase the profitability of banks. Thus, the business cycle affects net profit margin (through lending) and credit risk provision (through loan portfolio quality) (Ho Thi Hong Minh and Nguyen Thi Canh 2015). **Hypothesis 8: Economic growth (GDP) has a positive impact on the operational performance of commercial banks (H<sub>8</sub>).**

**INF – Inflation Rate:** The author uses the consumer price index CPI to measure the inflation rate:

$$\text{Inflationrate} = \frac{P_0 - P_{-1}}{P_{-1}} \times 100(\%)$$

In which:

- P<sub>0</sub>: average price level of the current period
- P<sub>-1</sub>: is the average price level of the previous period

Source: Delis (2012)

Bank as a special enterprise with business goods is currency, so the inflation rate directly affects the profitability of the bank. When inflation decreases, the purchasing power of Vietnam dong increases, at this time the price of gold and foreign currencies will decrease, so banks are convenient in mobilizing capital, lending and performing banking services. When inflation gets out of control, costs will increase, and at some point will destroy the entire economy. Revell (1979) said that the impact of inflation on the bank's business performance depends on the impact of inflation on salary and other operating



expenses of the bank. If the future inflation rate can be accurately forecasted, the bank can easily manage its operating expenses by adjusting the interest rate appropriately so that revenue grows faster than expenses and thereby earns more profit. According to Boyd et al. (2001), there is a significant negative economic relationship between inflation and banking sector development. Chirwa (2003); Syafri (2012); Adama and Apéléké (2017); Delis (2012) all believe that the inflation rate has a negative impact on the bank's operational performance. This makes it possible for inflation to contribute to the bank's financial performance and to engage in both interest and non-interest income activities. When the general price of goods increases, the operating costs of banks also increase, leading banks to increase their profit margins to compensate for the increase in operating costs. **Hypothesis 9: Inflation (INF) has a negative impact on the performance of commercial banks ( $H_9$ )** (Table 1).

### 3.3 Research Methods

Since most prior research was performed using a frequency approach, a priori information is not available. However, since the number of observations in the sample of 300 observations is relatively large, the priori information does not affect the posterior distribution too much. In this case, Block et al. (2011) proposed a standard Gaussian distribution with different a priori information (simulation of a priori information) and carried out Bayesian factor analysis to choose a simulation with the best priori news.

The simulations in Table 2 show decreasing levels of a priori information with Simulation 1 having the strongest priori information and Simulation 5 having the weakest priori information.

Similar to model 2, we also build 5 simulations (from simulation 6 to simulation 10) with simulation 6 having the strongest a priori information ( $\beta_i \sim N(0, 1)$ ) and model 10 having the strongest a priori information. Weakest a priori information ( $\beta_i \sim N(0, 10000)$ ).

In the next step, the research team carried out Bayesian regression for the above simulations, then performed Bayesian factor analysis (Bayes Factors) and Bayes test model (bayestest model). These are the techniques proposed by StataCorp LLC (2019) to select the simulation with the best a priori information. Basically, the Bayesian factor will provide a tool to compare the probability of a particular hypothesis (a priori information) to the probability of another hypothesis. It can be understood as a measure of the strength of evidence in favor of a theory among competing (information a priori) theories. Accordingly, Bayesian analysis will provide average Log BF (Bayes Factor – Bayesian factor), Log ML (Marginal Likelihood – marginal likelihood) and average DIC (Deviance Information Criterion – information bias); The posterior Bayesian test will help compare the posterior probability of the simulations with different a priori information, accordingly, based on the research data combined with the proposed a priori information, we will choose The simulation has the largest posterior probability  $P(M|y)$ .

In summary, in this study, the research team will build 5 simulations with 5 different a priori information, and Bayesian factor analysis and posterior Bayes test will help to choose a simulation with suitable a priori information. The simulation selected will be the one with the largest Log BF, Log ML average, minimum DIC mean and the largest  $P(M|y)$ .



**Table 1.** Expected signs of variables in the research model

Variable	Expected signs	Previous studies
Dependent variable		
ROA		Ho Thi Hong Minh and Nguyen Thi Canh (2015); Trujillo–Ponce (2013); Vincenzo Chiorazzo et al. (2008); Som Raj Nepali (2018); Le Long Hau and Pham Xuan Quynh (2017), Nguyen Minh Sang and Nguyen Thi Thuy Trang (2018)
Group of bank characteristics		
ICO <sub>NON</sub>	+	Apergis (2014); Saunders et al. (2014), Le Long Hau and Pham Xuan Quynh (2017), Nguyen Minh Sang and Nguyen Thi Thuy Trang (2018)
SIZE	+	Stiroh and Rumble (2006); DeYoung and Rice (2004b), Chiorazzo et al. (2008); Bourke, P. (1989); Adama and Apélété, T. (2017); Syafri (2012)
LOAN	+	Mercieca, Schaeck, and Wolfe (2007); Stiroh and Rumble (2006); Chiorazzo and authors (2008)
EQUITY	+	Abd Karim et al. (2010); Mercieca, Schaeck, and Wolfe (2007); Stiroh and Rumble (2006); Chiorazzo et al. (2008); Bourke (1989); Nguyen Tran Thinh (2013); Berger (1995)
COST	–	Bourke (1989); Guru et al. (2002); Syafri (2012)
DTL	+	Shiers (2002); Le Long Hau and Pham Xuan Quynh (2017); Ho Thi Hong Minh and Nguyen Thi Canh (2015)
LLP	–	Weersainghe and Perera (2013); Sufian and Habibullah (2009); Ahmad (2014); Nguyen Thanh Phong (2015)
Macro variables		
GDP	+	Obamuyi (2003); Adama and Apélété (2017); Chirwa (2003), Delis (2012)
INF	–	Chirwa (2003); Syafri (2012); Adama and Apélété (2017); Revell (1979), Delis (2012)

Source: Compiled by the author.

## 4 Results and Discussion

In Bayes Factor analysis, the selected model will be the one with the highest Log(BF), Log(ML) and the smallest DIC. The results in Table 3 show that model 1 with simulation 3 has the most advantage when Log (BF) and Log (ML) are the highest, however DIC is not as good as model 4 and 5. To ensure there is To be able to choose the most appropriate a priori information, the authors continue to analyze the Bayes test model, the results show that simulation 3 is the best when the P (Mly) criterion is superior to the other models remaining burn.

Bayesian analysis is simulated through the Markov Chain Monte Carlo (MCMC), therefore, to ensure the stability of the Bayesian regression, the MCMC series must

**Table 2.** Simulation of a priori information

Rational function	$ROA \sim N(\mu, \sigma)$
A priori distribution	
Simulation 1	$\alpha \sim N(0, 1)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 2	$\alpha \sim N(0, 10)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 3	$\alpha \sim N(0, 100)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 4	$\alpha \sim N(0, 1000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 5	$\alpha \sim N(0, 10000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$

Source: Compiled by the author.

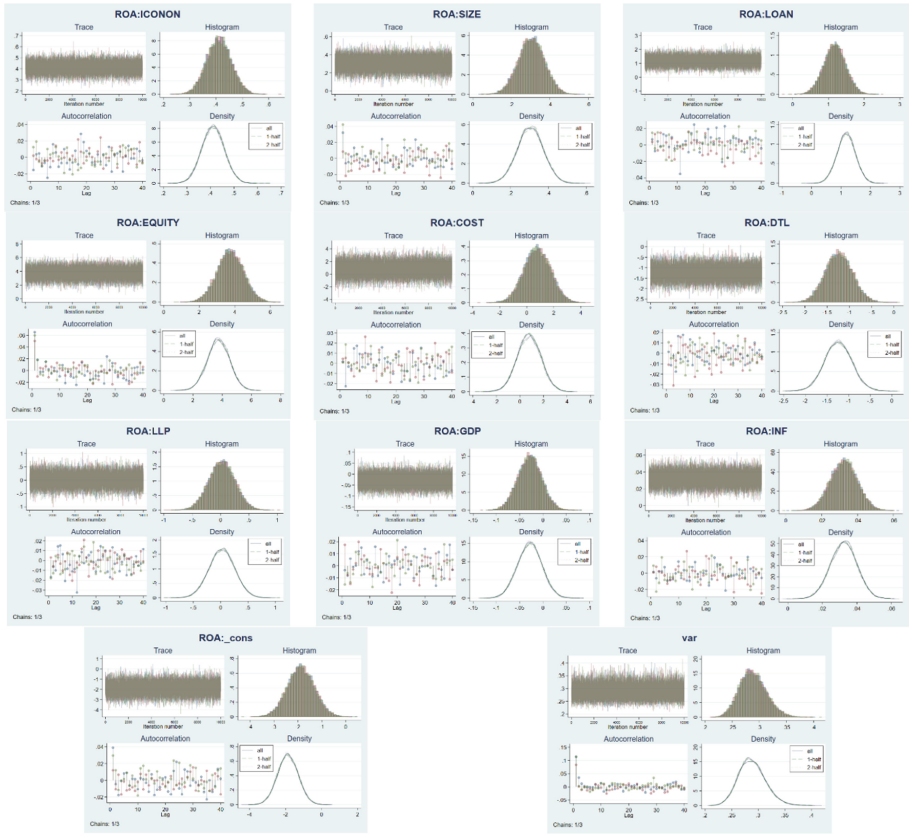
**Table 3.** Bayes Factor Analysis Results

	Chains	Avg DIC	Avg log (ML)	Log (BF)	P (Mly)
Simulation1	3	481.1574	-271.5035		0.000
Simulation2	3	445.5263	-258.7603	12.743	0.371
Simulation3	3	430.2282	-258.2345	13.269	0.628
Simulation4	3	427.6046	-265.2892	6.2143	0.001
Simulation5	3	427.6478	-276.0932	-4.589	0.000

Source: Compiled by the author.

converge, which means that the MCMC series must ensure stationarity. StataCorp LLC (2019) proposes that the MCMC series convergence test can be conducted through the convergence diagnostic graph.

According to StataCorp LLC (2019), the MCMC series convergence diagnostic graph includes trace plot, histogram, autocorrelation, and density plot. The trace plot helps to track the historical display of a parameter value over the iterations of the series, Fig. 1 shows the trace plot fluctuates around the mean value, so the MCMC series is stationary, that is, reaching convergence conditions. Besides, the autocorrelation chart in the graphs only fluctuates around the level below 0.02, according to StataCorp LLC (2019) the autocorrelation chart fluctuates around the level below 0.02, showing the agreement with the density the distribution and reflect all delays that are within the effective limit. According to StataCorp LLC (2019), the posterior distribution plot and density estimate show that the simulation of the shape of the normal distribution of the parameters, the histogram shape is uniform, it can be concluded that Bayes regression



**Fig. 1.** Convergence diagnostic graph Source: Calculations of the author

ensure stability. Thus, the results from Fig. 1 show that the MCMC series meets the convergence condition.

In addition to graphical convergence diagnostics, StataCorp LLC (2019) also recommends testing through Mean Acceptance Rate; Average minimum efficiency; and Gelman–Rubin  $R_c$  max. Table 4 shows that the model’s acceptance rate reaches 1, the model’s minimum efficiency is 0.912, far exceeding the allowable level of 0.01; In addition, the maximum  $R_c$  value of the coefficients is 1, Gelman and Rubin (1992) argue that the diagnostic value  $R_c$  of any coefficient of the model greater than 1.2 will be considered non–convergent. Thus, the values in Table 4 show that the MCMC series of the model satisfy the convergence requirements.

The regression results in Table 4 have determined that the variables DTL and GDP have a negative impact on ROA and the remaining variables help to improve ROA. Meanwhile, for the ROE variable, EQUITY, DTL, LPP and GDP have a negative effect while the variables ICONON, SIZE, LOAN and INF have a positive effect on ROE. Besides determining the sign of the regression coefficients, unlike the frequency method,

**Table 4.** Regression results

	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
ICONON	0.412	0.047	0.000	0.412	0.320	0.504
SIZE	0.304	0.068	0.000	0.305	0.170	0.436
LOAN	1.179	0.312	0.002	1.180	0.559	1.794
EQUITY	3.744	0.752	0.005	3.750	2.278	5.215
COST	0.787	0.996	0.006	0.787	-0.967	2.742
DTL	-1.240	0.311	0.002	-1.241	-1.853	-0.625
LLP	0.020	0.237	0.001	0.021	-0.444	0.482
GDP	-0.029	0.026	0.000	-0.029	-0.080	0.023
INF	0.032	0.008	0.000	0.032	0.017	0.047
_cons	-1.875	0.574	0.003	-1.881	-2.997	-0.737
var	0.288	0.025	0.000	0.287	0.244	0.342
Avg acceptance rate	1.000					
Avg efficiency min	0.811					
Max Gelman–Rubin Rc	1.000					

Source: Calculations of the author.

the Bayesian approach also allows us to calculate the probability of the occurrence of these effects (Table 5).

**Table 5.** Probabilistic test

ROA			
prob {ROA:ICONON} > 0	1.000	0.000	0.000
prob {ROA:SIZE} > 0	1.000	0.006	0.000
prob {ROA:LOAN} > 0	1.000	0.008	0.000
prob {ROA:EQUITY} > 0	1.000	0.000	0.000
prob {ROA:COST} > 0	0.896	0.410	0.002
prob {ROA:DTL} < 0	1.000	0.006	0.000
prob {ROA:LLP} > 0	0.536	0.499	0.003
prob {ROA:GDP} < 0	0.764	0.343	0.002
prob {ROA:INF} > 0	1.000	0.000	0.000

Source: Calculations of the author.

The variables  $ICO_{NON}$ , SIZE, LOAN, EQUITY, DTL, INF have a positive impact on the operational performance of banks (ROA) very clearly when the impact of these variables on ROA has a probability of touching 100%; The COST variable also tends to have a strong supportive effect on ROA when its probability of impact is approximately 90%. The variables LLP, GDP have a relatively weak impact on ROA when the overall probability is only 54% and 76%, respectively (Table 6).

**Table 6.** Comparison of expected results and research results of SMALL characteristic variables

Variable	Expectation	Result
$ICO_{NON}$	+	+
SIZE	+	+
LOAN	+	+
EQUITY	+	+
COST	–	+
DTL	+ /–	+
LLP	–	The impact is not clear

Source: *Compiled of author*

Note: + is the positive effect, – is the opposite effect

**Bank Size (SIZE):** The estimated results show that SIZE has a positive effect on ROA means that an increase in bank size will increase the bank's operational efficiency. This result coincides with the author's expectation. In previous empirical studies, this result is also consistent with the research results of authors such as Stiroh and Rumble (2006); DeYoung and Rice (2004b), Chiorazzo et al. (2008); Bourke (1989); Adama and Apélété (2017); Syafri (2012), etc. Vietnamese commercial banks with large scale and wide branches will have an advantage in capital mobilization, product and service development, and accessibility to customers. Higher banks, especially the competitiveness of large-scale banks will be stronger than those of small-sized banks, expanding the transaction network to central/populous areas to increase the number of customers. Therefore, an increase in bank size will increase profitability. In order to maximize the effectiveness of the network expansion and scale, commercial banks need to have specific plans to increase capital as well as improve the quality of products and services of the bank, thereby bringing benefits. Profits for commercial banks.

**Credit Size (LOAN):** The estimated results show that credit size has a positive impact on ROA. This result also coincides with previous empirical studies such as those of Mercieca, Schaeck, and Wolfe (2007); Stiroh and Rumble (2006); Chiorazzo et al. (2008). The research results reflect the reality of commercial banks in Vietnam that banks focus on lending to increase interest income and good loan quality will contribute to increasing bank profits. The bank's 3 main activities include providing credit, mobilizing deposits and performing payment intermediary functions. In particular, lending activities inherently account for a high proportion of income generation for banks. Therefore, when the

bank's lending rate increases, it will bring about a high rate of interest income (earnings from the difference between lending interest rates and deposit rates), thereby increasing the bank's profits.

**Equity Structure (EQUITY):** The estimation results show that capital structure has a positive effect on ROA. This result also coincides with previous experimental studies such as that of Abd Karim et al. (2010); Mercieca, Schaeck, and Wolfe (2007); Stiroh and Rumble (2006); Chiorazzo et al. (2008); Bourke (1989); Nguyen Tran Thinh (2013); Berger (1995). The increase in capital not only gives commercial banks the opportunity to expand credit, diversify products and services, improve financial capacity, and ensure financial ratios, but also protect customers. When the bank encounters risk during its operation, it helps the commercial bank to create a reputation among customers and investors. When the bank's financial capacity is improved, it will avoid wasting capital, save capital mobilization costs, and operate more effectively, so profitability will increase. When the equity ratio is high, the bank has a lower level of risk. Accordingly, banks with high capital ratios considered safer will be able to generate more profits. In addition, banks with a higher equity level often reduce the need for external capital, limiting the volatility of interest expenses, thereby positively affecting the bank's performance.

**Operating Cost (COST):** The estimated results show that capital structure has a positive impact on ROA. This does not coincide with the author's previous expectation, as well as the previous research results such as the study of Bourke (1989); Guru, Staunton, and Balashanmugam (2002); Syafri (2012). This can be explained that when banks begin to focus on increasing non-interest income ratio, increasing equity, investing in technology development, expanding branch network, transaction office services and increasing human resources to attract customers will incur additional operating costs. Thus, in case the bank's revenue increases, the higher the cost, the higher the profit of the bank will be.

**Deposit Size (DTL):** The estimated results show that deposit size has a positive effect on ROA. This result is consistent with studies of Shiers (2002); Le Long Hau and Pham Xuan Quynh (2017); Ho Thi Hong Minh and Nguyen Thi Canh (2015). Deposits are the main source of capital, accounting for a large proportion of mobilized capital in particular and business capital of commercial banks in general. The larger the deposit size, the greater the bank's ability to use capital, the more capital the bank has to finance lending activities, contributing to making the bank profitable. In fact, banks are currently competing with each other on deposit rates to attract depositors.

**Asset Quality (LLP):** Estimation results show that asset quality has no impact on ROA. This does not coincide with the author's previous expectations, as well as the results of previous studies such as that of Weersainghe and Perera (2013); Sufian and Habibullah (2009); Ahmad (2014); Nguyen Thanh Phong (2015). In other words, hypothesis H<sub>7</sub> is rejected.

**GDP Growth Rate (GROWTH):** The estimated results show that economic growth tends to have a negative impact on ROA. This does not coincide with the previous expectations of the author, as well as the results of previous studies such as the study of Obamuyi (2003); Adama and Apélété (2017); Chirwa (2003), Delis (2012).

**Inflation (INF):** The estimated results show that inflation has a positive impact on ROA with the probability of impact reaching the threshold of 100%. This does not coincide with the author's previous expectations, as well as the results of previous studies such as that of Chirwa (2003); Syafri (2012); Adama and Apélété (2017); Revell (1979), Delis (2012). This can be explained that when the inflation rate is high, commercial banks in Vietnam have policies to adjust interest rates accordingly raised in a timely manner, causing the difference between lending rates and deposit rates to increase, which will help increase the bank's profit. In addition, rising inflation reduces the real income of customers, thereby stimulating their investment and savings, so it will be easier for banks to raise capital with lower mobilization costs, help improve bank profitability.

## 5 Conclusion

### 5.1 Conclusion

Through significant statistical values with a confidence level of up to 99%, it can be affirmed that: developing in the direction of increasing the proportion of non-interest income generating activities will help commercial banks increase operational efficiency.

There are 07 factors that have a positive impact on the performance of Vietnamese commercial banks, including 06 internal factors which are non-interest income (ICONON), the size of the bank (SIZE), the size of the credit operation (LOAN), the capital structure (EQUITY), the operating cost (COST), the size of the deposit (DTL), and one external factor, in addition to the inflation rate (INF). Variables asset quality of banks (LLP) and economic growth rate (GDP) are not statistically significant.

### 5.2 Recommendations

**First, improving sales policies/models and diversifying products/services to improve service quality for customers:** Because they often have to face the constant competition of commercial banks, with both domestic and foreign trade, with the increasingly diverse needs and expectations of customers and the development of today's technology, the research into new products/services is indispensable in the operations of banks. Increasing the development of products such as cards, bond issuance, retail banking services, e-banking, etc. to create a distinction from your bank.

**Second, transfer some basic needs of existing customers to Kiosk Selfservices/Live Bank to reduce the load on the counter.** Currently on the market, there are only 3 banks that put the automatic banking system into operation, that is VP bank with VPBank NEO Express; TP Bank with Tien Phong Live Bank and Nam A Bank with Nam A OneBank digital ecosystem. Commercial banks need to build self-service transaction points (kiosk selfservices), which can provide most of the bank's products/services like traditional branches/transaction offices, but fully automated by modern machinery and technology. Accordingly, it is necessary to understand the customer's daily life journey, find out the customer's financial transaction needs such as frequency of use, when to use, how to choose a transaction location/method, etc.... Analyze customer's dissatisfaction/pain points during transaction at branch/transaction office, thereby assessing customer's need

to use selfservices/Live Bank kiosk and record customer's evaluations and suggestions for the Kiosk Selfservices/Live Bank model.

**Third**, *provide solutions to develop modern banking services, depending on each customer (individuals or businesses) to offer appropriate services.* Banks should encourage customers to use SMS banking, mobile app, internet banking, integrate QR payment on internet banking, link payment via e-wallets like Momo, Airpay, Moca, etc. non-cash payment method, encouraging customers to use the card by implementing attractive promotions (for example, the top 10 Visa cardholders with the highest total payment value of the week will receive 1 coupon code lucky to participate in the lucky draw program). This is also completely consistent with the development trend and orientation of commercial banks in Vietnam today, namely promoting banking activities in the direction of modernity and digital technology, similar to the orientation of Vietnam's commercial banks. Decision No. 1813/QĐ-TTg on the approval of the Project on development of non-cash payments in Vietnam for the period of 2021–2025 issued by the Prime Minister of Vietnam on October 28, 2021.

**Fourth**, *increase non-interest income from commissions and fees by promoting cross-selling of products and services to existing customers, stimulating demand and selling more products and services that customers have not used, especially focus on exploiting the group of customers who have loans at the bank.*

**Fifth**, *regularly train human resources.* Since non-profit activities often use modern technology, employees must be knowledgeable and able to use technology proficiently. Commercial banks need to regularly organize training courses to improve their professional skills as well as how to use new banking technology to improve the professional qualifications of their staff. High-quality human resources will help the bank bring products/services to customers more easily and conveniently, increasing customer satisfaction and willingness to recommend the bank to relatives, friends, colleagues. In addition, it is necessary to pay attention to the quality of governance and internal inspection and control. This work must be regularly updated and upgraded in parallel with the development of technology. In addition, banks also need to improve the image, customer service and professionalism of their staff so that customers have the best experience when using the bank's products and services, thereby building a strong brand, reputation to gain trust from customers.

**Sixth**, *there is a specific strategy to develop non-interest income:* Although non-interest income has a positive impact on the performance of commercial banks, credit and savings are still the main activities of commercial banks. Therefore, commercial banks need to build the most appropriate, optimal ratio of non-interest income to total income in the direction of decreasing dependence on traditional activities. Specifically, in the structure of non-interest income, the bank should have a ratio for each type such as income from services; income from business, investment, etc. to have reasonable promotion solutions.



## References

### English

- Abd Karim, M.Z., Chan, S.G., Hassan, S.: Bank efficiency and non-performing loans: evidence from Malaysia and Singapore. *Prague Economic Papers*, vol. 2, no. 1 (2010)
- Abdul, L.A.: Income diversification and bank efficiency in an emerging market. *Manag. Financ.* **41**(12), 1318–1335 (2015)
- Adama, C. Apélété, T.: The bank sector performance and macroeconomics environment: empirical evidence in Togo. *Int. J. Econ. Finance* **9**(2) (2017)
- Ahmad, A.A.: Impact of internal factors on bank profitability: comparative study between Saudi Arabia and Jordan. *J. Appl. Finance Bank.* **4**(1), 125–140 (2014)
- Al-Muharrami, S., Matthews, K.: Market power versus efficient-structure in Arab GCC banking. *Cardiff Economics Working Papers* (2009)
- Al-Tarawneh, A., Abu Khalaf, B., Al Assaf, G.: Noninterest income and financial performance at Jordanian Banks. *Int. J. Financ. Res.* **8**(1), 166–171 (2017)
- Apergis, N.: The long-term role of non-traditional banking in profitability and risk profiles: evidence from a panel of U.S. banking institutions. *J. Int. Money* **45**, 61–73 (2014)
- Atony, F., Ludger, S., Vito, T.: Public sector efficiency: evidence for new EU member states and emerging markets. *Working Paper Series*, No. 581, p. 9 (2006)
- Baele, L., Jonghe, O.D., Vennet, R.V.: Does the stock market value bank diversification? *J. Bank. Finance* **3**(1), 1999–2023 (2007)
- Bailey-Tapper, S.A.: Non-interest Income, Financial Performance & The Macroeconomy: Evidence on Jamaican Panel Data. *Bank of Jamaica (BOJ) Working Paper* (2010)
- Bain, J.S.: Relation of profit rate to industry concentration American manufacturing, 1936–1940. *Q. J. Econ.* **65**, 293–324 (1951)
- Berger, A., Hasan, I., Zhou, M.: The effects of focus versus diversification on bank performance: evidence from Chinese banks. *J. Bank. Finance* **34**, 1417–1435 (2010)
- Berger, A.N.: The profit-structure relationship in banking: test of market-power and efficient-structure hypotheses. *J. Money Credit Bank.* **27**, 404–431 (1995)
- Bourke, P.: Concentration and other determinants of bank profitability in Europe, North America and Australia. *J. Bank. Finance* **13**(1), 65–79 (1989)
- Boyd, J.H., Levine, R., Smith, B.D.: The impact of inflation on financial sector Performance. *J. Monet. Econ.* **47**, 221–248 (2001)
- Busch, R., Kick, T.K.: Income diversification in the German banking industry (2009)
- Chang, H.H., Boisvert, R.N., Hung, L.Y.: Land subsidence, production efficiency, and the decision of aquacultural firms in Taiwan to discontinue production. *Ecol. Econ.* **69**(12), 2448–2456 (2010)
- Chiorazzo, V., Milani, C., Salvini, F.: Income diversification and bank performance: evidence from Italian banks. *J. Financ. Serv. Res.* **33**(3), 181–203 (2008)
- Chirwa, E.W.: Determinants of commercial bank's profitability in Malawi: a cointegration approach. *Appl. Financ. Econ.* **13**(8), 565–571 (2003)
- Choudhry, M.: *An Introduction to Banking: Principles, Strategy and risk Management*. Wiley (2018)
- Chronopoulos, D.K., Girardone, C., Nankervis, J.C.: Are there any cost and profit efficiency gains in financial conglomeration? Evidence from the accession countries. *Eur. J. Finance* **17**(8), 603–621 (2011)
- Craigwell, R., Maxwell, C.: Non-Interest Income at Commercial Banks in Barbados: An Empirical Note. *Central Bank of Barbados, Mimeo*, October 2005

- Delis, M.D.: Bank competition, financial reform, and institutions: the importance of being developed. *J. Dev. Econ.* **97**(2), 450–465 (2012)
- DeYoung, R., Rice, T.: How do banks make money? The facilities of fee income. *Economic Perspectives*, Federal Reserve Bank of Chicago (2004a)
- DeYoung, R., Rice, T.: Non-interest income and financial performance at U.S.A commercial bank. *Financ. Rev.* **39**(1), 456–478 (2004b)
- Elsas, R., Hackethal, A., Holzhauser, M.: The anatomy of diversification. *J. Bank. Finance* **34**(6), 1274–1287 (2010)
- Elyasiani, E., Wang, Y.: Bank holding company diversification and production efficiency. *Appl. Financ. Econ.* **22**(17), 1409–1428 (2012)
- Elyasiani, E., Mehdiian, S.M.: A non-parametric approach to measurement of efficiency and technological change: the case of large US commercial banks. *J. Financ. Serv. Res.* **4**, 157–168 (1990a)
- Elyasiani, E., Mehdiian, S.M.: Efficiency in the commercial banking industry: a production frontier approach. *Appl. Econ.* **2**, 539–551 (1990b)
- Farrell, M.J.: The measurement of productive efficiency. *J. Roy. Stat. Soc.. Ser. A (Gen.)* **120**(3), 253–290 (1957)
- Gamra, S.B., Plihon, D.: Revenue diversification in emerging market banks: implications for financial performance. arXiv preprint [arXiv:1107.0170](https://arxiv.org/abs/1107.0170) (2011)
- Getter, D.E.: Overview of Commercial Bank (Depository) Banking and Industry Conditions. Congressional Research Service (2016)
- Guru, B.K., Staunton, J., Balashanmugam, B.: Determinants of commercial bank profitability in Malaysia. *J. Money Credit Bank.* **17**(1), 69–82 (2002)
- Klein, P.G., Saldenberg, M.R.: *Diversification, Organization, and Efficiency: Evidence from Bank Holding Companies*, pp. 97–127. Wharton School Center for Financial Institutions, University of Pennsylvania (1997)
- Koponen, T.M.: Commodities in action: measuring embeddedness and imposing values. *Sociol. Rev.* **50**(4), 543–569 (2003)
- Lepetit, L., Nys, E., Rous, P., Tarazi, A.: Bank income structure and risk: an empirical analysis of European banks. *J. Bank. Finance* **32**, 1452–1467 (2008)
- Li, F., Zou, Y.: The impact of credit risk management on profitability of commercial banks. A study of Europe. Thesis. Umea School of Business and Economics (2014)
- Li, L., Zhang, Y.: Are there diversification benefits of increasing noninterest income in the Chinese banking industry? *J. Empir. Finance* **24**, 151–165 (2013)
- Lozano-Vivas, A., Pasiouras, F.: The impact of non-traditional activities on the estimation of bank efficiency: international evidence. *J. Bank. Finance* **34**(7), 1436–1449 (2010)
- Markowitz, H.: Portfolio selection. *J. Finance* **7**(1), 77–91 (1952)
- Mercieca, S., Schaeck, K., Wolfe, S.: Small European banks: benefits from diversification? *J. Bank. Finance* **31**, 1975–1998 (2007)
- Meslier, C., Tacneng, R., Tarazi, A.: Is bank income diversification beneficial? Evidence from an emerging economy. *J. Int. Financ. Markets. Inst. Money* **31**, 97–126 (2014)
- Mester, L.J.: Efficient production of financial services: scale and scope economies. *Bus. Rev.* 15–25 (1987)
- Nepali, S.R.: Income diversification and bank risk-return trade-off on the Nepalese commercial banks. *Asian Econ. Financ. Rev.* **8**(2), 279–293 (2018)
- Obamuyi Marshal, T.: Determinant of bank's profitability in a developing economy: evidence for Nigeria. *Organiz. Mark. Emerg. Econ.* **4**(2), 97–111 (2013)
- Odesanmi, S., Wolfe, S.: Revenue diversification and insolvency risk: evidence from banks in emerging economies. *Soc. Sci. Res. Netw.* (2007)
- Oniang'o, R.: Effect of non-interest income on profitability of commercial banks in Kenya (Doctoral dissertation, University of Nairobi) (2015)

- Revell, J.: Inflation and Financial Institution. *Financial Times*, London (1979)
- Saunders, A., Schmid, M., Walter, I.: Non-interest income and bank performance: does ring-fencing reduce bank risk? (Working Paper No. 1417) (2014)
- Shiers, A.: Branch banking, economic diversity and bank risk. *Q. Rev. Econ. Finance* **42**, 587–598 (2002)
- Singh, K.B., Upadhyay, Y., Singh, S.K., Singh, A.: Impact of non-interest income on risk and profitability of banks in India. In: *Make in India: A Road Ahead*, pp. 997–1007 (2016)
- Soedarmono, W., Machrouh, F., Tarazi, A.: Bank market power, economic growth and financial stability: evidence from Asian banks. *J. Asian Econ.* **22**(6), 460–470 (2011)
- Stiroh, K.: Diversification in banking: is noninterest income the answer? *J. Money Credit Bank.* **36**, 853–882 (2004)
- Stiroh, K.J., Rumble, A.: The dark side of diversification: the case of US financial holding companies. *J. Bank. Finance* **30**(8), 2131–2161 (2006)
- Sufian, F.: Profitability of the Korean banking sector: panel evidence on bank-specific and macroeconomic determinants. *J. Econ. Manag.* **7**(1), 43–72 (2011)
- Sufian, F., Habibullah, M.S.: Bank specific and macroeconomic determinants of bank profitability: empirical evidence from the China banking sector. *Front. Econ. China* **4**(2), 274–291 (2009)
- Syafri, M.: Factors affecting bank profitability in Indonesia. In: *The 2012 International Conference on Business and Management*, pp. 236–242 (2012)
- Tarazi, A., Crouzille, C., Tacneng, R.: Bank Diversification, Risk and Profitability in an Emerging Economy with Regulatory Asset Structure Constraints: Evidence from the Philippines, 22 February 2010 (2010)
- Tariq, W., Usman, M., Mir, H.Z., Aman, I., Ali, I.: Determinants of commercial banks profitability: empirical evidence from Pakistan. *Int. J. Account. Financ. Report.* **4**(2), 1–22 (2014)
- Trujillo-Ponce, A.: What determines the profitability of banks? Evidence from Spain. *Account. Finance* **53**(2), 561–586 (2013)
- Tuyishime, R., Mamba, F., Mbera, Z.: The effects of deposits mobilization on financial performance in commercial banks in Rwanda. A case of Equity Bank Rwanda Limited. *Int. J. Small Bus. Entrep. Res.* **3**(6), 44–71 (2015)
- Weersainghe, V.E.I.W., Perera, T.R.: Determinants of profitability of commercial banks in Sri Lanka. *Int. J. Arts Commer.* **2**(10), 141–170 (2013)
- Williams, B.: The impact of non-interest income on bank risk in Australia. *J. Bank. Finance* **73**, 16–37 (2016)

## Vietnamese

- Bộ Y tế – Bộ Tài Chính: Thông tư liên tịch Số: 37/2015/TTLT–BYT–BTC quy định thống nhất giá dịch vụ khám bệnh, chữa bệnh bảo hiểm y tế giữa các bệnh viện cùng hạng trên toàn quốc (2015)
- Chính phủ: Quyết định số 254/QĐ–TTg của Thủ tướng Chính phủ phê duyệt “Đề án cơ cấu lại hệ thống các tổ chức tín dụng giai đoạn 2011–2015”, Hà Nội (2012)
- Chính phủ: Quyết định số 986/QĐ–TTg của Thủ tướng Chính phủ phê duyệt “Chiến lược phát triển ngành ngân hàng Việt Nam đến năm 2025, định hướng đến năm 2030”, Hà Nội (2018)
- Tiến, H.N., Hiền, V.T.: Trao đổi về phương pháp tính tỷ lệ thu nhập ngoài tín dụng của ngân hàng thương mại. *Công nghệ Ngân hàng* **48**, 36–39 (2010)
- Minh, H.T.H., Canh, N.T.: Đa dạng hoá thu nhập và các yếu tố tác động đến khả năng sinh lời của các ngân hàng thương mại Việt Nam. *Công nghệ Ngân hàng* **106**(107), 13–23 (2015)
- Lê, L.H., Phạm, X.Q.: Tác động của đa dạng hoá thu nhập đến hiệu quả kinh doanh của các ngân hàng thương mại Việt Nam. *Công nghệ Ngân hàng* **124**, 11–22 (2016)

- Ngân hàng Nhà nước Việt Nam: Thông tư Số: 41/2016/TT-NHNN quy định tỷ lệ an toàn vốn đối với ngân hàng, chi nhánh ngân hàng nước ngoài (2016)
- Ngân hàng Nhà nước Việt Nam: Văn bản hợp nhất 20/VBHN-NHNN hợp nhất thông tư quy định về cấp giấy phép và tổ chức, hoạt động của ngân hàng thương mại, chi nhánh ngân hàng nước ngoài, văn phòng đại diện của tổ chức tín dụng nước ngoài, tổ chức nước ngoài khác có hoạt động ngân hàng tại Việt Nam do ngân hàng nhà nước Việt Nam ban hành (2018)
- Nguyễn, K.M.: Từ điển toán kinh tế, thống kê, kinh tế lượng Anh-Việt. Nhà xuất bản Khoa học và kỹ thuật (2004)
- Nguyễn, M.K.: Nghiệp vụ Ngân Hàng Thương Mại. Nhà xuất bản Thống kê (2009)
- Nguyễn, M.S.: Tác động của đa dạng hoá thu nhập đến hiệu quả hoạt động của các ngân hàng thương mại tại Việt Nam. Kinh tế và Phát triển 241, 40-49 (2017)
- Sáng, N.M., Nguyễn, T.T.T.: Tác động của thu nhập ngoài lãi đến rủi ro và khả năng sinh lời của các ngân hàng thương mại Việt Nam, Tạp chí Khoa học Đại học Đà Lạt 8(1S), 118-132 (2018)
- Phong, N.T.: Phân tích các yếu tố ảnh hưởng đến lợi nhuận của các ngân hàng thương mại niêm yết trên thị trường chứng khoán Việt Nam. Luận văn Thạc sĩ, Trường Đại học Tài chính - Marketing thành phố Hồ Chí Minh (2015)
- Hiền, N.T.D., Hạp, N.H.: Thu nhập ngoài lãi và hiệu quả tài chính tại các ngân hàng thương mại Việt Nam. Tạp chí Công nghệ Ngân hàng, số 127, tháng 10/2016 (2016)
- Thịnh, T.N.: Phân tích các yếu tố tác động đến lợi nhuận các Ngân hàng niêm yết trên thị trường chứng khoán Việt Nam. Luận văn Thạc sĩ, Trường Đại học Kinh tế TP.HCM (2013)
- Peter, S.R.: Quản trị Ngân hàng Thương mại, Nhà xuất bản Tài chính (2004)
- Trình, P.T.T.: Kinh tế lượng ứng dụng trong kinh tế và tài chính. Nhà xuất bản kinh tế TP. Hồ Chí Minh (2016)
- Quốc hội nước Cộng hòa xã hội chủ nghĩa Việt Nam khóa XII: Luật các tổ chức tín dụng, 47/2010/QH12, Hà Nội (2010)
- Thủ Tướng Chính Phủ: Quyết định Số: 1813/QĐ-TTg về việc phê duyệt đề án phát triển thanh toán không dùng tiền mặt tại Việt Nam giai đoạn 2021-2025 (2021)
- Trịnh, T.T.H., Nguyễn, H.P., Lê, T.T.: Tác động của đa dạng hóa thu nhập đến hiệu quả hoạt động của các ngân hàng thương mại Việt Nam. Tạp chí tài chính 25, 10-29 (2018)
- Vinh, V.X., Mai, T.T.P.: Lợi nhuận và rủi ro từ đa dạng hoá thu nhập của ngân hàng thương mại Việt Nam. Tạp chí Phát triển kinh tế 26(8), 54-70 (2015)



# Cryptocurrency Portfolio Management Based on Usage Characteristics Criteria Applying R-Vine Copula

Terdthiti Chitkasame<sup>1</sup>, Pichayakone Rakpho<sup>2</sup>, and Nachattapong Kaewsompong<sup>1</sup>(✉)

<sup>1</sup> Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand  
terdthiti\_c@cmu.ac.th, nachat\_flysky@hotmail.com

<sup>2</sup> Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University,  
Chiang Mai, Thailand

**Abstract.** This study proposes cryptocurrency portfolio management with asset selection based on usage criteria, as the direction of price variations between usage categories varies according to conventional financial concepts. The study used data from seven major types of cryptocurrencies: the storage of value, Smart contracts, Decentralized finance, Value transfer, Oracle, and Stable coins and the Meme type. The data is collected from a daily time series for the period from 20 November 2019 to 21 March 2022. The study used the Vine Copula method to analyze dependence before estimating the optimal portfolio with the least variance based on Markowitz's theory. These results imply that the least variance portfolio should incorporate ETH, BUSD, and BTC. It further suggests that the portfolio include some DiFi. In addition, Doge coin (Meme type) is the root node of all cryptocurrency assets. When Doge coin volatility is high, investors should be cautious.

**Keywords:** Cryptocurrency · Portfolio management · R-Vine Copula

## 1 Introduction

Global money and finance are undergoing dramatic changes. Digitized assets and novel financial channels, tools, and systems are reshaping financial transaction paradigms and establishing alternative capital conduits [40]. Cryptocurrency is a digital or virtual currency asset that is based on a network that is distributed across a large number of computers. It is a type of digital asset that is based on a network of computers. After Satoshi invented Bitcoin in 2008 [28], many new cryptocurrencies spurred the creation known as Altcoins (Alternative coins) with the aim of solving the Bitcoin constraints problems. Thus, this result is the cause of many cryptocurrencies being produced. These cryptocurrency innovations and the perceived investment potential have led to a rapid growth in the number of altcoins and the market size of cryptocurrency [10]. According to Coin Market Capital, it has approximately 869 cryptocurrencies currently trading around the world with a combined market capitalization of US\$2.05 trillion circulating supply at the end of March. There are increased about 9.04% year-on-year and the total

crypto market volume over the last 24 h on April 1, 2022, is US\$135.96 billion which makes it a 23.82% increase.

While an increasing number of retailers globally accept cryptocurrency as an asset of investment or saving, some conservative investors believe it is only a passing fad, with no intrinsic value and maybe the perfect vehicle for forming a bubble in investments [35]. This is due to a number of issues, including price volatility [21]. For instance, one Bitcoin cost \$6,600. (July 2018). The price dropped significantly in mid-December 2018 - to \$3,200. Then, at the end of June 2019, the price increased dramatically - to \$13,000. [18]. As can be seen, the price of cryptocurrency assets has fluctuated significantly over time, despite the price consistently increasing over the long term. However, it cannot be denied that investors face price uncertainty during the holding period. This means that the total risks will increase.

Accordingly, investors and traders in the Cryptocurrency market are eager to refine their knowledge of the determinants of volatility for risk reduction and portfolio management [42]. Choosing a basket (portfolio) of crypto assets to invest in is essential, given that these assets are characterized by strong price swings that can lead to reliance shifts and portfolio gains or losses (see, e.g., [3, 8, 13, 27]). Previous cryptocurrency portfolio studies, like Bouri [6], Corbet [11], Kajtazi and Moro [20], and Platanakis and Urquhart [30] have attempted to include crypto-assets, particularly Bitcoin, into conventional portfolios. Almost all claim significant gains from incorporating Bitcoin into regular strategies. While there have also been numerous studies that have attempted to do only cryptocurrency portfolios for instance the study of Smales [34], and Platanakis and Urquhart [30]. They focused on a different method, but they both choose assets based on the same criteria (Market shares). However, the similarity of the assets that meet the selection criteria presents a problem. This indicated that the assets belonged to the same category or possess comparable utilization characteristics, and consequently their asset prices may move in the same way. Therefore, using these criteria to select assets for portfolios based on conventional finance theories may not be acceptable. This research provides a solution to the aforementioned challenges pertaining to the selection of cryptocurrency assets where this research refers to Finnomena and Bitkub's research based on classification criteria according to usage characteristics. They have separated the cryptocurrency into seven-main types, consisting of the Store of value, the Smart contract, Decentralized finance, the Value transfer, the Oracle, the Stable coin, and the Meme type. Afterwards, we chose representatives for each asset type based on their highest market value. The conclusion of this study will be an optimal portfolio based on Markowitz's theory (1991) by applied the Vine-copula approach.

Portfolio management firstly is required to capture the asset volatility. Then, the generalized autoregressive conditional heteroskedasticity (GARCH) model is applied mostly to capture the price volatility. While Zhu and Galbraith's [44] study faced the finding of cryptocurrency risk may be incorrect in inference by the original GARCH model because it is also a financial asset, and probably has asymmetric, leptokurtic, and is heavy tailed in data. Hence, the EGARCH and GJR-GARCH models were developed to address this issue. These models are capable of capturing asymmetry effects referring to the negative unexpected returns that have a greater impact on future volatility than positive unexpected returns and the leverage effect, which refers to the correlation

between shocks and subsequent shocks to volatility [39]. These characteristics have an effect on projections and on the effectiveness of the Value at Risk (VaR) and the Expected shortfall.

The discussion continues regarding the most effective way to quantify the dependence structure between cryptocurrencies. Copula analysis is frequently used in risk management and asset pricing. The copula is a multivariate distribution function that can be used to describe the non-linear dependence structure between each variable, especially financial assets. It creates a slew of novel metrics that allow us to fully exploit dependence in some extreme conditions, such as when in a financial crisis.

However, the types of multivariate copula become progressively limited, and it automatically assigns the same dependence structure to each pair of margins when the dimensions of the data increases [9]. It then fails to capture the dependence structure variability between pairs of margins [25]. In response to these shortcomings, Trucios et al. [38] have designed a recent multivariate copula technique that is dubbed the vine-copula. The vine-copula method is capable of independently capturing each bivariate dependence structure by constructing a multi-level tree [1].

Therefore, our paper attempts to examine the price dependency relationships and portfolio optimization of seven cryptocurrencies of data, which represent the seven main types of cryptocurrencies. These consist of the Store of value types represented by Bitcoin (BTC) [22], the Smart contract type is represented by as Ethereum (ETH), the Defi type is represented by as Dai (DAI), the Value transfer type is represented by as Ripple (XRP), the Oracle type is represented by as Chainlink (LINK), the Stable coin type is represented by as Binance USD (BUSD), and the Meme type is represented by as Doge coin (Doge) using the vine copulas with the GARCH(1, 1)-types model to allow for the modelling of composite risks of financial assets. As more investors and institutions including banks, hedge funds, and even governments join the burgeoning cryptocurrency ecosystem it is crucial to identify technical tools that can profitably trade these crypto assets despite their inherent volatility thereby choosing the best judgments while constructing portfolios in order to minimize risks.

The following sections outline the research methods used in this work. The second section discusses the methodology utilized in this investigation. Section 3 contains the data sources and an explanation of the data. Section 4 discusses the empirical findings. Finally, Sect. 5 summarizes some key findings and conclusions.

## 2 Methodology

### 2.1 GARCH-Type Models

#### 2.1.1 GARCH Model

Bollerslev [5] proposed the GARCH model. GARCH (1, 1) is taken into account in this study and can be expressed as follows:

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \varepsilon_t &= \sigma_t \cdot v_t, \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \end{aligned} \tag{1}$$

where  $\omega, \alpha, \beta \geq 0$  are unknown parameters with parameter restrictions  $\alpha + \beta \leq 1$ .  $\varepsilon_t$  is the uncorrelated random variable with mean zero and variance  $\sigma_t^2$ .  $v_t$  is a standardized residual of a chosen innovation.

**2.1.2 EGARCH Model**

Nelson [29] proposed EGARCH, a nonlinear GARCH model, to capture the long-memory and short-memory volatility effects, as well as the asymmetric leverage effects, of financial variables. The EGARCH (1, 1) model’s conditional variance specification is

$$\ln(\sigma_t^2) = \omega + \gamma \left| \frac{\varepsilon_{t-1}}{\sqrt{\sigma_{t-1}^2}} \right| + \alpha \left( \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right) + \beta \ln(\sigma_{t-1}^2), \tag{2}$$

where  $\gamma$  is the asymmetric leverage coefficient to describe the volatility leverage effect.

**2.1.3 GJR-GARCH Model**

Glosten et al. [14] proposed the GJR-GARCH model. The GJR-GARCH model is another popular nonlinear GARCH model. It is used to capture the potential larger impact of negative shocks on data volatility (asymmetric leverage volatility effect). The GJR-GARCH (1, 1) model is defined as follows:

$$\sigma_t^2 = \omega + (\alpha + \gamma I_{t-1}) \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{3}$$

$$I_{t-1} = \begin{cases} 0 & \text{if } \varepsilon_{t-1} \geq 0, \\ 1 & \text{if } \varepsilon_{t-1} < 0. \end{cases} \tag{4}$$

and  $\gamma$  is an asymmetric leverage effect.

**2.2 Copula**

In this study, we will begin by describing Copula briefly. Copula was proposed by Sklar [33]. Copulas are functions that connect two random variables using multivariate distribution functions. According to Sklar’s theorem, any multivariate distribution can be factored into marginal cumulative distributions. It can be written as follows:

$$F(Y) = C(F_1(y_1), \dots, F_d(y_d)), \forall Y = (y_1, \dots, y_d), \tag{5}$$

where  $C(u_1, \dots, u_d)$  is a cumulative distribution function (cdf) with uniform marginals on the unit interval. If  $F_i(y_i)$  is the CDF of a univariate continuous random variable  $y_i$ . Then  $C(F_1(y_1), \dots, F_d(y_d))$  is a d-variate distribution for  $Y = (y_1, \dots, y_d)$  with marginal distributions  $F_i$ . The corresponding density is as follows:

$$C(F_1, (y_1), \dots, F_d(y_d)) = \frac{h(F_1^{(-1)}(u_1), \dots, F_d^{(-1)}(u_d))}{\prod_{i=1}^d f_i(F_i^{(-1)}(u_i))}, \tag{6}$$

where  $h$  is the density function associated to  $H$ ,  $f_i$  is the density function of each marginal distribution and  $C$  is the copula density.



### 2.3 Vine Copula

It was proposed by Aas et al. [1] defined as any pair-copula construction. In the study, we illustrate the pair-copula construction for seven dimensions. Consider seven random variables  $x = x_1, x_2, x_3, \dots, x_7 \sim F$  with marginal distribution functions  $F_1, F_2, F_3, \dots, F_7$  and density functions is  $f_1, f_2, f_3, \dots, f_7$ .

#### 2.3.1 C-Vines Copula Is Formula Below

$$f(x) = \prod_{i=1}^d f(x_i) \prod_{j=1}^{d-1} \prod_{h=1}^{d-j} c_{j,j+h|1,2,\dots,j-1} (F(x_j | x_{1,j-1}), F(x_{j+h} | x_{1,j-1})) \tag{7}$$

#### 2.3.2 D-Vines Copula Is Formula Below

$$f(x) = \prod_{i=1}^d f(x_i) \prod_{j=1}^{d-1} \prod_{h=1}^{d-j} c_{h,h+j|h+1,\dots,h+j-1} (F(x_h | x_{h-1:h+j-1}), F(x_{h+j} | x_{h-1:h+j-1})) \tag{8}$$

#### 2.3.3 R-Vines Copula

Forasmuch as the C-Vine pair copula lacks flexibility when dealing with complicated models, Dissman, [12] proposes an alternative by constructing an R-vine using a diagram algorithm. The R-vine has a few flaws, one of which is its static nature. Furthermore, as the dimension increases, the computational effort required to estimate the model grows exponentially. The following is a general description of an R-vine:

$$f_{1,\dots,d}(x) = \prod_{k=1}^d f_k(x_k) \prod_{i=1}^{d-1} \prod_{e \in E_i} C_{C_{e,a}, C_{e,a} | D_e} (F_{C_{e,a} | D_e}(x_{C_{e,a}} | x_{D_e}), F_{C_{e,b} | D_e}(x_{C_{e,a}} | x_{D_e})) \tag{9}$$

where  $x = (x_1, \dots, x_d)$ ,  $e = \{a, b\}$ ,  $xx_k = \sum_{t=1}^k (x_t - \bar{x})$ ,  $yy_k = \sum_{t=1}^k (y_t - \bar{y})$ ,  $k = 1, 2, \dots, N$  are the variables of  $D_e$ ,  $f_i$  is the inverse function of  $F_i (i = 1, \dots, n)$ .

### 2.4 Value-at-Risk (VaR)

Value at Risk (VaR) and conditioned Value at Risk or Expected Shortfall (ES) has been widely used to measure risk since the 1990s. The VaR of portfolio can be written as

$$VaR_\alpha = \inf \{l \in R : P(L > l) \leq 1 - \alpha\} \tag{10}$$

where,  $\alpha$  is a confidence level with a value [0,1] which presents the probability of Loss  $L$  to exceed  $l$  but not larger than  $1 - \alpha$ . While an alternative method, ES, is the extension of the VaR approach to remedy two conceptual problems of VaR [15]. Firstly, VaR measures only percentiles of profit-loss distribution with difficulty to control for non-normal distribution. Secondly, VaR is not sub-additive. ES can be written as

$$ES_{\alpha} = E(L|L > VaR_{\alpha}) \tag{11}$$

To find the optimal portfolios, Rockafellar and Uryasev [41] introduced portfolio optimization by calculating VaR and extend VaR to optimize ES. The approach focuses on the minimizing of ES to obtain the optimal weight of a large number of instruments. In other words, we can write the problem as in the following [43]

$$\text{Minimize } ES_{\alpha} = E(L|L > \inf\{l \in R : P(L < l) \leq 1 - \alpha\}) \tag{12}$$

The objective function is to

$$\begin{aligned}
 R_p &= \sum_{i=1}^n (w_i r_i) \\
 &\sum_{i=1}^n (w_i) \\
 0 &\leq w_i \leq 1, i = 1, 2, \dots, n
 \end{aligned}
 \tag{13}$$

where  $R_p$  is an expected return of the portfolios,  $W_i$  is a vector of weight portfolio, and  $r_i$  is the return of each instrument.

### 2.5 Portfolio Optimization

In this section, we analysis of the optimal investment portfolio in cryptocurrencies. There are two methods based on portfolio optimization are imposed in terms of the equally weighted portfolio and the global minimum variance portfolio.

## 3 Data Specification

In this study indicates the model using seven cryptocurrencies data which represent the seven main types of cryptocurrencies, consisting of the Store of value type which is represented by as Bitcoin (BTC), the Smart contract type is represented by as Ethereum (ETH), the Defi type is represented by as Dai (DAI), the Value transfer type is represented by as Ripple (XRP), the Oracle type is represented by as Chainlink (LINK), the Stablecoin type is represented by as Binance USD (BUSD), and the Meme type is represented by as Dogecoin (DOGE). The data is gathered from the daily time series for the period from 20 November 2019 to 21 March 2022. All data for this study was collected from <https://charts.coinmetrics.io/network-data>. Additionally, each price of cryptocurrency was transformed into a return by the formula log return  $r_t = \log(P_t - P_{t-1})$ . The summary statistics are displayed in Table 1. The six cryptocurrency returns are positive average return rates except for DAI return which exhibits a negative average return rate. In the Jarque-Bera normality test, the result found that all of the returns data are not normally distributed (see MBF of Jarque-Bera test are zero). In addition, in the Augmented Dickey-Fuller test (ADF) unit root test, the result found that all return cryptocurrencies are the null hypothesis. This means that all returns are stationary (see MBF Unit root test are zero).

**Table 1.** Descriptive Statistics

	BTC	ETH	DAI	XRP	LINK	BUSD	DOGE
Mean	0.002	0.003	-3.20E-06	0.001	0.002	3.0E-06	0.004
Median	0.002	0.004	-3.20E-06	0.001	0.003	-1.40E-05	0.000
Maximum	0.167	0.244	0.051	0.423	0.258	0.015	1.407
Minimum	-0.47	-0.565	-0.049	-0.521	-0.636	-0.008	-0.513
Std. Dev	0.040	0.052	0.003	0.065	0.067	0.001	0.092
Skewness	-1.784	-1.569	0.022	-0.052	-1.095	3.598	5.204
Kurtosis	26.160	20.831	83.374	15.764	14.161	71.533	74.795
Jarque-Bera	19518.41	11650.94	2219.60	5791.11	4600.31	168772.21	187045.1
MBF of Jarque-Bera	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unit root test	-31.533	-31.684	-22.427	-29.529	-31.724	-21.994	-15.311
MBF Unit root test	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	853	853	853	853	853	853	853

Note: MBF is the Minimum Bayes factor.

## 4 Estimated Results

### 4.1 Selected Models

In this study, we used three-different classes of vine copula (C-vine, D-vine, and R-vine), and the GARCH (1, 1)-type model was used as a marginal model. It included the original GARCH, EGARCH, and GJR-GARCH models with normal, student-t, and skew-student-t distributions. The cryptocurrency returns are considered here for modeling the dependency of the seven main types of cryptocurrencies. Then, the minimum value of AIC was employed as an information criterion for selecting the most suitable models. The estimated results of model selections is presented in Table 2. The results indicate that the best-fit model is the R-vine copula based on the EGARCH (1, 1) and Student-t distributions.

**Table 2.** AIC for selected models

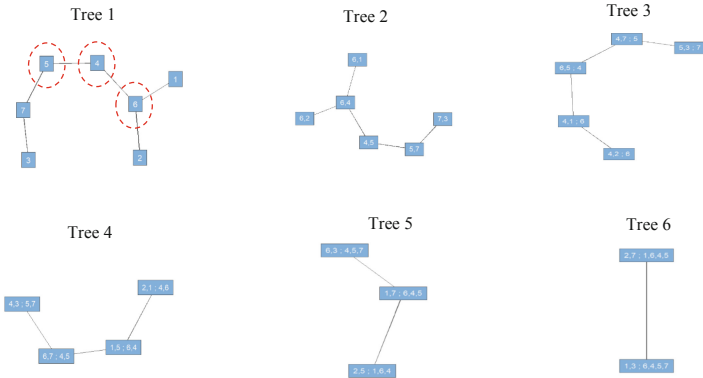
Models		Normal	Student-t	Skew-Student-t
C-Vine	GARCH	13.251	12.884	13.258
	EGARCH	9.472	-1.755	8.254
	GJR-GARCH	4.843	8.831	9.185
D-Vine	GARCH	12.141	11.142	14.242
	EGARCH	8.482	4.414	10.274
	GJR-GARCH	3.843	7.271	2.314
R-Vine	GARCH	11.711	11.774	12.084
	EGARCH	8.481	<b>-2.854</b>	10.442
	GJR-GARCH	3.75	7.631	2.12

### 4.2 Estimation of the R-Vine Copula EGARCH (1, 1) Model

Note: The data namely 1: BTC, 2: ETH, 3:XRP, 4:LINK, 5:BUSD, 6:DOGE, and 7:DAI. Family copula namely Joe270: Rotate Joe copula (270°), Gumbel: Gumbel copula, sClayton: survival Clayton copula, Gaussian: Gaussian copula, sGumbel: survival Gumbel copula, Clayton90: Rotate Clayton copula (90°), sJoe: survival Joe copula, and Frank: Frank copula. Parameters namely :Kendall’s tau, :upper tail dependence coefficient, :lower tail dependence coefficient.

In this section, we analyzed the tree structure formed by the seven main types of cryptocurrencies using the R-vine copula based on the EGARCH (1, 1) model which consisted of 6-tree presented in Table 2. The dependence structure between the seven main types of cryptocurrencies of the R-vine is captured by six types of bivariate copulas: Joe270, Gumbel, sClayton, Gaussian, sGumbel, and Clayton90.

Additionally, the range of Kendall’s  $\tau$  in tree  $[-0.05, 0.07]$ . The indices of upper and lower tail dependence on the tree structure are mostly equal to 0, except for the pair of BUSD and DAI which was upper tail dependent equal to 0.05, except for the pair of LINK and BUSD of which the upper tail dependence was equal to 0.01, and the pair of DOGE and ETH which is lower tail dependence was equal to 0.04. It is clear that LINK, BUSD, and DOGE dominate the dependency structure across the seven major types of cryptocurrencies, implying that they serve as the primary stress transmitters between cryptocurrencies as shown in Tree 1 of Fig. 1. This result provides the Doge coin as the root node, and this means that the Doge coin was the center of volatility of effect to each asset (Table 3).



**Fig. 1.** The tree structure of the R-vine copula. Note: The number present 1: BTC, 2: ETH, 3: XRP, 4: LINK, 5: BUSD, 6: DOGE, and 7: DAI

**Table 3.** The R-Vine copula EGARCH (1, 1) model

Tree 1	Margins	7,3	5,7	4,5	6,1	6,2	6,4
	Family	Joe270	Gumbel	sClayton	Gaussian	sGumbel	Clayton90
	$\tau$	-0.03	0.041	0.072	-0.051	0.032	-0.011
	$\lambda_U$	0	0.051	0.01	0	0	0
	$\lambda_L$	0	0	0	0	0.04	0
Tree 2	Margins	5,3  7	4,7  5	6,5  4	4,1  6	4,2  6	
	Family	Clayton90	sClayton	sJoe	Clayton90	Frank	
	$\tau$	-0.03	0.02	0.02	-0.03	0.02	
	$\lambda_U$	0	0	0	0	0	
	$\lambda_L$	0	0	0.05	0	0	
Tree 3	Margins	4,3 5,7	6,7 4,5	1,5 6,4	2,1 4,6		
	Family	Gaussian	Frank	sClayton	Frank		
	$\tau$	0.043	-0.011	0.031	0.022		
	$\lambda_U$	0	0	0	0		
	$\lambda_L$	0	0	0	0		
Tree 4	Margins	6,3 4,5,7	1,7 6,4,5	2,5 1,6,4			
	Family	Joe90	Gaussian	Clayton90			
	$\tau$	-0.001	0.002	-0.01			
	$\lambda_U$	0	0	0			
	$\lambda_L$	0	0	0			
Tree 5	Margins	1,3 6,4,5,7	2,7 1,6,4,5				
	Family	sClayton	Gaussian				
	$\tau$	0.01	0.02				
	$\lambda_U$	0	0				
	$\lambda_L$	0	0				
Tree 6	Margins	2,3 1,6,4,5,7					
	Family	sClayton					
	$\tau$	-0.02					
	$\lambda_U$	0					
	$\lambda_L$	0					

### 4.3 Portfolio Optimization

Evaluation of the performance of optimized portfolios is in this section. By comparing the risk-based portfolios' risk (as ES) and return to an equally weighted portfolio as a benchmark then focused on the minimum variance portfolio which was calculated by the model presented in Table 4. It reported the optimal investment proportion of the cryptocurrency's portfolio of the expected shortfall was at 95% confidence interval. The result confirmed that this portfolio had a lower return by about 0.02 %, but it had

**Table 4.** Optimal investment proportion of cryptocurrencies portfolio

N = 1,000	ES 5%								
	BTC	ETH	XRP	LINK	BUSD	DOGE	DAI	Return	Risk
Equally weighted portfolio	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	49.80%	11.03%
Minimum variance portfolio	0.1539	0.1572	0.1477	0.1321	0.1556	0.1409	0.1127	49.78%	10.97%

Note: N: number of simulated data

a lower risk than the benchmark (about 0.06%). This result illustrated that the optimal weights suggest that investors should focus on minimum variance, which in this study the risk was equal to 10.97 % which was smaller than the equally weighted portfolio (the benchmark), and it had returned equal to 49.78%.

These results indicate that the best portfolio should contain a high proportion of ETH, followed by BUSD and BTC, consistent with the fact that BUSD is the lowest asset because its value is pegged to the 1 US dollar (by a Binance algorithm), and BTC still has the highest proportion. This result accords with the studies cited in this research ([6, 11, 20], and [30]). BTC had significant gains and a positive impact on the entire return. However, the BUSD has the second-highest weight in our suggested portfolio, corresponding with the findings of Tenkam et al. [37], as the stable coin can mitigate or hedge against risk.

In short, this research implemented R-vine copulas to determine the dependence structure between cryptocurrencies. The copulas selected using the R-vine copula framework are more sensitive to the asset’s return tail distributions and asymmetries of the series, providing the R-vine copula approach more flexibility and better outcomes than the typical bivariate copula framework. Finally, we employ R-vine copula’s dependencies for portfolio optimization. This process is consistent with the findings of Boako et al. [4], Trucios et al. [38], Tenkam et al. [37], and Mba et al. [24]. Comparing R-vine copulas to C-vine copulas, D-vine copulas, and multivariate copulas, they discovered that R-vine copulas is more flexible and suitable for cryptocurrencies’ returns. Furthermore, we compared the returns and risks of our portfolio to those of previous research. We find that our result is similar and slightly better in terms of portfolio returns being slightly higher while the risks are slightly lower. These outcomes may be a result of the different times of data.

## 5 Conclusion

The goal of this study had been to propose a cryptocurrency portfolio management with asset selection based on usage of criteria of the direction of price fluctuations between the different usage categories that varied according to the conventional financial concepts. The data used in the study consisted of seven cryptocurrencies of data, which represent the seven main types of cryptocurrencies. This was Value storage, Smart contracts, Decentralized finance, Value transfer, Oracle, Stable coins, and Meme types represented by as BTC, ETH, DAI, XRP, LINK, BUSD, and Doge, respectively. This study applied the Vine Copula method to investigate the dependence structure before estimating the optimal portfolio with minimum variance based on Markowitz’s theory.

These results suggest that the minimum variance portfolio should contain a high proportion of ETH, followed by BUSD and BTC. It further suggests that the portfolio should contain a small amount of DiFi type. In addition to the Vine Copula result, the Doge coin (Meme type) is the root node of all cryptocurrency assets. Investors, including institutional investors, should take precautions when the Dogecoin's volatility is significant.

**Acknowledgements.** This research work was partially supported by Chiang Mai University, Thailand.

## References

1. Aas, K., Czado, C., Frigessi, A., Bakken, H.: Pair-copula constructions of multiple dependence. *Insur.: Math. Econ.* **44**(2), 182–198 (2009)
2. Artzner, P.: Thinking coherently. *Risk*, 68–71 (1997). Bedford, T., Cooke, R.M.: Vines—a new graphical model for dependent random variables. *Ann. Stat.* **30**(4), 1031–1068 (2002)
3. Bekiros, S., Hernandez, J.A., Hammoudeh, S., Nguyen, D.K.: Multivariate dependence risk and portfolio optimization: an application to mining stock portfolios. *Resour. Policy* **46**, 1–11 (2015)
4. Boako, G., Tiwari, A.K., Roubaud, D.: Vine copula-based dependence and portfolio value-at-risk analysis of the cryptocurrency market. *Int. Econ.* **158**, 77–90 (2019)
5. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **31**(3), 307–327 (1986)
6. Bouri, E., Molnár, P., Azzi, G., Roubaud, D., Hagfors, L.I.: On the hedge and safe haven properties of bitcoin: is it really more than a diversifier? *Financ. Res. Lett.* **20**, 192–198 (2017)
7. Brauneis, A., Mestel, R.: Cryptocurrency-portfolios in a mean-variance framework. *Financ. Res. Lett.* **28**, 259–264 (2018)
8. Brunnermeier, M.K., Pedersen, L.H.: Market liquidity and funding liquidity. *Rev. Financ. Stud.* **22**(6), 2201–2238 (2009)
9. Brechmann, E.C., Czado, C.: Risk management with high-dimensional vine copulas: an analysis of the Euro Stoxx 50. *Stat. Risk Model.* **30**(4), 307–342 (2013)
10. Chuen, D.L.K., Guo, L., Wang, Y.: Cryptocurrency: a new investment opportunity? *J. Alternat. Invest.* **20**(3), 16–40 (2017)
11. Corbet, S., Meegan, A., Larkin, C., Lucey, B., Yarovaya, L.: Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Econom. Lett.* **165**, 28–34 (2018)
12. Dissman, J.F.: Statistical inference for regular vines and application. Thesis, Technische Universität München, Zentrum Mathematik (2010)
13. Florackis, C., Kontonikas, A., Kostakis, A.: Stock market liquidity and macro-liquidity shocks: evidence from the 2007–2009 financial crisis. *J. Int. Money Financ.* **44**, 97–117 (2014)
14. Glosten, L.R., Jagannathan, R., Runkle, D.E.: On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Financ.* **48**(5), 1779–1801 (1993)
15. Halulu, S.: Quantifying the risk of portfolios containing stocks and commodities. Doctoral dissertation, Bogazici University (2012)
16. Hileman, G., Rauchs, M.: Global cryptocurrency benchmarking study. *Camb. Cent. Alternative Financ.* **33**, 33–113 (2017)
17. Hyunyoung, C., Hal, V.: Predicting the present with google trends. *Econ. Rec.* **88**(s1), 2–9 (2012)

18. Hrytsiuk, P., Babych, T.: The cryptocurrencies risk measure based on the Laplace distribution. In: M3E2-MLPEED, pp. 261–276 (2020)
19. Hrytsiuk, P., Babych, T., Bachyshyna, L.: Cryptocurrency portfolio optimization using value-at-risk measure. *Adv. Econ. Bus. Manage. Res.* **95**, 385–389 (2019)
20. Kajtazi, A., Moro, A.: The role of bitcoin in well diversified portfolios: a comparative global study. *Int. Rev. Financ. Anal.* **61**, 143–157 (2019)
21. Katsiampa, P.: Volatility estimation for bitcoin: a comparison of GARCH models. *Econ. Lett.* **158**, 3–6 (2017)
22. Kubat, M.: Virtual currency bitcoin in the scope of money definition and store of value. *Procedia Econ. Financ.* **30**, 409–416 (2015)
23. Markowitz, H.M.: Foundations of portfolio theory. *J. Financ.* **46**(2), 469–477 (1991)
24. Mba, J.C., Mwambi, S.: A Markov-switching COGARCH approach to cryptocurrency portfolio selection and optimization. *Fin. Mark. Portfolio Manage.* **34**(2), 199–214 (2020). <https://doi.org/10.1007/s11408-020-00346-4>
25. McNeil, A.J., Nešlehová, J.: Multivariate Archimedean copulas, d-monotone functions and  $l_1$ -norm symmetric distributions. *Ann. Stat.* **37**(5B), 3059–3097 (2009)
26. Morgan, J.P.: RiskMetrics Technical Document. JP Morgan, New York (1996)
27. Moshirian, F.: The global financial crisis and the evolution of markets, institutions and regulation. *J. Bank. Financ.* **35**(3), 502–511 (2011)
28. Nakamoto, S., Bitcoin, A.: A peer-to-peer electronic cash system. *Bitcoin* 4, 2 (2008). <https://bitcoin.org/bitcoin.pdf>
29. Nelson, B.D.: Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347–70 (1991)
30. Platanakis, E., Urquhart, A.: Should investors include bitcoin in their portfolios? A portfolio theory approach (2018). <http://ssrn.com/abstract=3215321>
31. Platanakis, E., Urquhart, A.: Should investors include bitcoin in their portfolios? A portfolio theory approach. *Br. Account. Rev.* **52**(4), 100837 (2020)
32. Schuh, S.D., Shy, O.: U.S. consumers' adoption and use of bitcoin and other virtual currencies. Unpublished; slides of preliminary findings (2016). <https://payments.nacha.org/sites/payments.nacha.org/files/files/Virtual%20Currency.pdf>. Accessed 20 Mar 2017
33. Sklar, A.: Random variables, joint distribution functions, and copulas. *Kybernetika* **9**(6), 449–460 (1973)
34. Smales, L.A.: Bitcoin as a safe haven: is it even worth considering? *Financ. Res. Lett.* **30**, 385–393 (2019)
35. Tasca, P., Liu, S., Hayes, A.: The evolution of the bitcoin economy: extracting and analyzing the network of payment relationships (2016). Accessed 20 Mar 2017
36. Tasca, P., Tessone, C.J.: Taxonomy of blockchain technologies. Principles of identification and classification. arXiv preprint [arXiv:1708.04872](https://arxiv.org/abs/1708.04872) (2017)
37. Tenkam, H.M., Mba, J.C., Mwambi, S.M.: Optimization and diversification of cryptocurrency portfolios: a composite copula-based approach. *Appl. Sci.* **12**(13), 6408 (2022)
38. Trucios, C., Tiwari, A.K., Alqahtani, F.: Value-at-risk and expected shortfall in cryptocurrencies' portfolio: a vine copula-based approach. *Appl. Econ.* **52**(24), 2580–2593 (2020)
39. Rakpho, P., Yamaka, W., Phadkantha, R.: Predicting energy price volatility using hybrid artificial neural networks with GARCH-type models. In: Honda, K., Entani, T., Ubukata, S., Huynh, V.N., Inuiguchi, M. (eds.) IUKM 2022. LNCS, vol. 13199, pp. 317–328. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-98018-4\\_26](https://doi.org/10.1007/978-3-030-98018-4_26)
40. Rauchs, M., Hileman, G.: Global cryptocurrency benchmarking study. *Camb. Cent. Alternative Financ. Rep.* (2017)
41. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. *J. Bank. Financ.* **26**(7), 1443–1471 (1959)



42. Walther, T., Klein, T., Bouri, E.: Exogenous drivers of bitcoin and cryptocurrency volatility—a mixed data sampling approach to forecasting. *J. Int. Finan. Markets. Inst. Money* **63**, 101133 (2019)
43. Yang, B., Tarkhamtham, P., Phuensan, P., Zhu, K.: Portfolios optimization under regime switching model: evidences in the American bonds and other financial assets. In: Sriboonchitta, S., Kreinovich, V., Yamaka, W. (eds.) *Behavioral Predictive Modeling in Economics. Studies in Computational Intelligence*, vol. 897, pp. 377–392. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-49728-6\\_25](https://doi.org/10.1007/978-3-030-49728-6_25)
44. Zhu, D., Galbraith, J.W.: A generalized asymmetric student-t distribution with application to financial econometrics. *J. Econom.* **157**(2), 297–305 (2010)



# Does Debt Affect Profitability of Construction Companies in Vietnam? A Bayesian Approach

Bui Dan Thanh<sup>(✉)</sup> and Nguyen Ngoc Huyen<sup>(✉)</sup>

HCMC University of Banking, Ho Chi Minh City, Vietnam  
tanhbhd@buh.edu.vn, huyen6801@gmail.com

**Abstract.** This study aims to find out the effects of debt on profitability of construction companies listed on the Stock Exchange in Vietnam in the period from 2010 to 2020. Using secondary data from 72 enterprises listing construction, including 792 observations with Bayesian regression technique to find out the factors affecting the overall debt ratio of enterprises. Regression results show that there are 6 important factors affecting the profitability ratio, namely short-term debt to total assets (SDA), long-term debt to total assets (LDA), liquidity (LQ), annual revenue growth (GROWTH), firm size (SIZE), inflation rate (INF). From there, business managers can refer to the research results to make decisions in the course of business operations, ensuring compliance with the development goals of enterprises in the construction industry.

**Keywords:** Capital structure · construction firms · Bayesian regression

## 1 Introduction

The construction industry is one of the key economic sectors of Vietnam. In the past 15 years, foreign companies have not stopped pouring investment capital into Vietnam. Typically, in 2018, FDI capital increased by 9.1% compared to 2017 (according to data from the Foreign Investment Department - Ministry of Planning and Investment). High-rise buildings such as Bitexco Financial Tower, Keangnam Landmark Tower, Vincom Landmark 81 are growing up more and more. Accordingly, Vietnam's construction industry has also made great progress. However, the stronger the commercialization and integration process, the more challenges construction businesses face such as capital management, competitors, decline of building materials, etc. leading to suboptimal business performance.

The main reason affecting the business performance of construction enterprises is that the debt capital is often very large, especially in the construction industry, which occupies capital for a long period of time. Therefore, determining the influence of debt in order to use debt effectively will help Vietnamese construction enterprises achieve high profit margins.

Determining the capital structure of Vietnamese construction enterprises is the main goal of this study. In particular, we define the optimum level of debt utilized to finance the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 248–263, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_17](https://doi.org/10.1007/978-3-031-35763-3_17)

operations of construction enterprises in growing countries like Vietnam. We anticipate that the results of this study will help financial managers in the future make the best decisions about capital structure policies.

## 2 Literature Review

### 2.1 The Modern Theory of Capital Structure (M&M Theory)

#### **Capital Structure Has no Effect on Firm Value (M&M 1958)**

In 1958, Modigliani and Miller (M&M) investigated whether the cost of capital increased or decreased as a firm increased or decreased its external debt. M&M made several assumptions about perfect capital markets: No transaction costs; all investors can borrow or lend at the same interest rate; no bankruptcy costs and financial distress costs; assuming the risk class is homogeneous, i.e. businesses operating under similar conditions will have the same level of business risk; no income tax.

If capital markets are perfect, M&M assumes that businesses that do the same business and expect the same annual returns should have the same value, regardless of capital structure because the value of the business must depend on its activities rather than depending on the form of funding. From this it can be concluded that all firms with the same expected returns and the same value should have the same average cost of capital at all levels of debt and equity ratios.

Although the assumptions of perfect capital markets are unrealistic, there are two assumptions that need to be emphasized because these assumptions have an impact on the results of M&M's research:

- No taxation: This is an important issue and one of the key advantages of debt is the effect of tax shields;
- M&M's theoretical risk is calculated entirely by the volatility of cash flows. M&M ignores the possibility that cash flows may stop because of default. This is a significant problem compared to other theories if debt is high.

These assumptions bring only one advantage (debt is cheaper and less risky for investors) and one disadvantage of borrowing money (cost of equity increases with debt ratio because of debt ratio to total capital).

Thus, according to the point of view of M&M (1958), in a perfect market, the value of an unlevered firm is also equal to the value of a debt enterprise, or in other words, the value of a firm is independent of its capital structure.

#### **Capital Structure Affects the Value of the Firm (Modigliani and Miller 1963)**

In 1963, M&M launched a follow-up study with the elimination of the corporate income tax hypothesis. According to M&M, with corporate income tax, the use of debt will increase the value of the business. Since interest expense is a reasonable expense that is deductible when calculating income tax, a portion of the income of a debt-using business is passed on to investors, resulting in the value of the debt-using business equal to value of the non-debt firm plus the gain from the use of debt.

Thus, according to the M&M tax model (1963), capital structure is related to the value of the firm. The higher the use of debt, the higher the value of the business increases and increases until the business is financed with 100% debt.

## 2.2 Pecking Order Theory

The pecking order theory originates from the research of Donaldson (1961). This study provides evidence that many executives prefer to use internal financing and only consider external financing (debt and issue of new shares) in cases where the capital needs are inevitable increase. This theory was further investigated by Myers and Majluf (1984). They argue that corporate financing decisions are based on asymmetric information. Asymmetric information is a phrase that indicates that CFOs know their company's value better than outside investors. Asymmetric information affects the choice between internal and external funding; between new issuance of debt securities and equity securities.

## 2.3 Trade – off Theory

The trade-off theory of capital structure is based on M&M theory, which considers the costs of financial distress and the effects of taxes. According to research by Kraus and Litzenberger (1973) and Myers (1977), in contrast to M&M theory, the value of a firm should only accept a specific level of debt to maximize firm value. According to the trade-off hypothesis, the target capital structure is the extent to which the benefits of the tax shield can offset the costs of financial distress. However, the cost of financial distress will outweigh the benefit of the interest tax shield when the debt ratio reaches a specific point. The value of the company will then decrease, increasing the possibility of bankruptcy.

From a capital structure trade-off point of view, the following factors have an impact on capital structure: corporate income tax, costs associated with financial distress, tangible fixed assets, firm size and profitability.

The last part of the M&M theory on the financial crisis costs of financially distressed firms has been addressed by the trade-off theory. However, there are also many things that trade-off theory cannot explain, such as why some businesses continue to succeed, good business results with very little debt, or the fact that a company has more likely to issue shares when stock prices are high and the company needs external funding (instead of debt).

## 2.4 Comprehensive Study

Abor (2005) conducted a study on 22 companies listed on the Ghana exchange in the period 1998–2002. The results show that the ratio of short-term debt to total assets (STD) has a significant positive relationship with ROE. According to the author's argument, the company using short-term debt will cost less, increasing profits. In contrast, the results show a significant negative relationship of long-term debt to total assets (LTD) with ROE. An increase in long-term debt that is associated with a decrease in profitability has a higher cost of long-term debt and vice versa. The research results also show a positive

relationship between the debt-to-total assets (TD) ratio and the efficiency of corporate financial management. This shows that an increase in debt is associated with an increase in profits. Therefore, the company with higher debt will have a higher profit. And the results also show that the profits of the companies in the sample also increase due to the increase in the size and revenue of the company.

Tian and Zeitun (2007) conducted a study based on data of 167 companies for the period 1989–2003. The study aimed to examine the influence of capital structure on firm performance in Jordan. The research results show that the ratio of short-term debt to total assets has a positive impact on the performance of the business, specifically, companies with a high ratio of short-term debt to total assets will have a high growth and profitability performance.

Ahmad et al. (2012) studied the relationship between capital structure and performance among firms, using data from the reports of 58 companies listed on the Malaysian stock market in the period 2005–2010. According to research, operational efficiency is measured by ROE and ROA. The research results show that the variables: the ratio of short-term debt to total assets, the ratio of long-term debt to total assets, the ratio of total debt to total assets, the growth rate of assets and business performance have the positive effect on ROE. Besides, the results also show that the variables of long-term debt to total assets and revenue growth rate have no significant influence on corporate profits.

Pouraghajan and Malekian (2012) analyze the impact of capital structure on the financial performance of companies listed on the Tehran Stock Exchange. For this purpose, they studied a sample of 400 companies in the form of 12 industrial groups between 2006 and 2010. In the study, the variables ROA and ROE were used to measure the financial performance of the companies. The results show that there is a significant negative relationship between debt ratio and financial performance of companies and a significant positive relationship between asset turnover, firm size, tangible assets ratio and growth opportunities with measures of financial performance. In addition, the research results show that by reducing the debt ratio, management can increase the profitability of the company.

Bui Dan Thanh (2016) conducted a study using data of 1032 small and medium-sized enterprises in Ho Chi Minh City in the period 2006–2014. The results show that the ratio of total debt to total assets and the ratio of short-term debt to total assets has a positive relationship with the financial performance (ROA, ROE) of the enterprise. In addition, the control variables: firm size and the ratio of fixed assets to total assets have a negative impact on ROA, ROE.

Doan Ngoc Phuc (2018) studied the impact of capital structure on the performance of 217 companies listed on the 2 Stock Exchanges of Ho Chi Minh City and Hanoi in the period 2007–2012. Research results shows that long-term debt has a positive impact on ROA and ROE, while short-term debt and total debt have a statistically significant negative impact on firm performance measured by ROA and ROE.

Nguyen Thi Dieu Chi (2018) used Tobit model to conduct a study of 116 service enterprises listed on the Vietnamese stock market in the period 2010–2018 on the impact of debt capital structure on financial results. Research results show that both short-term debt structure and long-term debt structure have a negative impact on corporate profitability. Specifically, when an enterprise uses too much debt in its capital structure,

it will have negative effects on financial performance, or the more debt the enterprise borrows, the lower the financial performance of the enterprise, due to increased costs from loan interest.

It is noteworthy that the aforementioned research used frequency approaches or descriptive analyses with suitably large sample sizes to analyze capital structure in sample enterprises. Based on a dataset of 72 enterprises listing construction on the Vietnam Stock Exchange in the period from 2010 to 2020, this study used Bayesian logistic regression with informative priors. The research has made the following contributions, as expected: (i) Help financial managers in the future make the best decisions about capital structure policies.; (ii) By using Bayesian MCMC simulations in informative (thoughtful) prior settings, our findings enable a generalized conclusion that, in contrast to frequentist approaches, Bayesian estimation using thoughtful priors can provide meaningful results.

### 3 Model and Data

#### 3.1 General Model

$$Y_{it} = \beta_0 + \sum \beta_i X_{it} + u_{it}$$

In which:  $i$ : the  $i$ -th cross unit (data of one or more variables collected for multiple sample units or sample locations at the same time) and  $t$  is the  $t$ -th time;  $Y_{it}$  is the dependent variable;  $X_{it}$  is the independent variable;  $\alpha$ : coefficient of freedom,  $\beta$ : coefficient of regression,  $u_{it}$ : residual.

#### 3.2 Research Model

Based on empirical studies in the world and in Vietnam, the author found that the number of variables as well as the way to measure the variable and the results of the direction of the impact of the variables on the profitability ratio is different in different research. However, these studies all selected some of the effects of debt on corporate profitability such as the ratio of short-term debt to total assets, the ratio of long-term debt to total assets, liquidity, revenue growth rate, business size and inflation rate. These variables all have the ability to collect data and all have economic significance, are correlated and explain the research problem. Therefore, the author has built the research model of the topic as follows:

$$ROA_{it} = a + b_1 SDA_{it} + b_2 LDA_{it} + b_3 LQ_{it} + b_4 GROWTH_{it} + b_5 SIZE_{it} + b_6 INF_{it} + u_{it}$$

In there: - ROA: Dependent variable in observation time  $i$  in period  $t$ . Return on total assets, representing the profit rate of the business.

- SDA: Ratio of short-term debt to total assets; LDA: Ratio of long-term debt to total assets; LQ: Liquidity; GROWTH: Revenue growth rate; SIZE: Logarithm of Total assets; INF: Inflation rate.

### 3.3 Variables and Hypotheses

From the review of previous studies, the author proposes the following research hypotheses:

#### Dependent Variable

**Return on Total Assets – ROA:** Return on total assets is measured by profit after tax on total assets. It is a pure measure of a business's efficiency in generating returns on assets that are not affected by management's funding decisions. According to the studies of Gleason et al. (2000), Tian and Zeitun (2007), Ahmad et al. (2012) all chose ROA as a measure of the profitability of the business. The research results show that the higher the ROA of the enterprise, the better the capital investment efficiency of the enterprise.

#### Independent Variables

**Short-Term Debt-to-Total Assets Ratio – SDA:** The ratio of short-term debt to total assets indicates how much of a percentage of total capital the company uses short-term debt to finance its assets. According to Zeitun and Tian (2007), SDA is calculated according to the formula:

$$SDA = \frac{\text{Current Liabilities}}{\text{Total Assets}}$$

According to the research results of Abor (2005), Gill (2011), Bui Dan Thanh (2016) show that SDA has a positive relationship with the profitability of enterprises. In contrast, the research results of Tian and Zeitun (2007), Ahmad et al. (2012) suggest that SDA has a negative relationship with the profitability of the business. The peculiarity of construction enterprises is that the use of short-term debt is very large, accounting for a small proportion of total assets. This can reduce profitability and bring burden on businesses. This runs counter to the optimal capital structure theory, which traditionally holds that the cost of capital can be reduced by increasing the use of debt. Thus, the higher the ratio of short-term debt to total assets, the lower the profit of the business. ***Hypothesis H1: The short-term debt ratio (SDA) has a negative effect on the profitability of the business.***

**Long-Term Debt to Total Assets Ratio – LDA:** The ratio of long-term debt to total assets shows how much of a business's long-term debt is used to finance its assets. According to Abor (2005), long-term debt ratio (LDA) is calculated according to the formula:

$$LDA = \frac{\text{Long term Debt}}{\text{Total Assets}}$$

Large capital occupation and prolonged business cycle lead to increased interest costs from long-term borrowing of construction enterprises. This can have a negative impact on financial performance, reducing corporate profits. In fact, the Vietnamese economy in particular and the world economy in general are constantly fluctuating. If the economy has bad changes, construction enterprises will have difficulty in getting loans. According to the research results of Abor (2005), Tian and Zeitun (2007) also show that, the higher the LDA, the lower the corporate profit. ***Hypothesis H2: Long-term debt ratio (LDA) has a negative effect on corporate profitability.***

**Liquidity – LQ:** Liquidity is the ability to convert into cash to pay for short-term debts of the business. According to Githaiga and Karibu (2015), liquidity is measured as follows:

$$LQ = \frac{\text{Current Assets}}{\text{Current Liabilities}}$$

In the total capital of construction enterprises in Vietnam, short-term capital accounts for a higher proportion than long-term capital. This limits long-term investment capital. At that time, the enterprise must increase an additional cost to maintain liquidity. The liquidity of each short-term debt is also different. According to Goddard et al. (2004), Molyneux and Thornton (1992) argue that a higher level of liquidity also creates an opportunity cost due to lower returns compared to other assets. *Hypothesis H3: Liquidity (LQ) has a negative effect on the profitability of the business.*

**Revenue Growth Rate – GROWTH:** Revenue growth rate shows the relative revenue growth (in %) over time. Thereby, see the business situation of the enterprise. According to Tian and Zeitun (2007), the revenue growth rate is calculated as follows:

$$GROWTH = \frac{\text{Net Revenue}_n - \text{Net Revenue}_{n-1}}{\text{Net Revenue}_{n-1}} \times 100\%$$

In there n: years. The studies of Zeitun and Tian (2007), Nunes et al. (2009) all show that revenue growth rate has a positive relationship with business performance. Myers (1977) also pointed out that: Firms with high revenue growth have more options for future investment than firms with low revenue growth. Myers' comments are almost consistent with construction industry businesses. The specificity of the industry is to create products of great value, so the profit from business activities is often very high. This helps businesses have good revenue growth and attract many investors. *Hypothesis H4: Revenue growth rate (GROWTH) has a positive effect on the profit margin of the business.*

**Firm Size – SIZE:** According to Abor (2005) and Ahmad (2012), enterprise size is calculated as follows:

$$SIZE = \text{Log}_e(\text{Total Assets})$$

Construction corporations and enterprises often raise capital by issuing shares on the stock market or borrowing from credit institutions. The larger the size of the business, the more opportunities it has. Because a large-scale business will easily make customers have more confidence and often receive more incentives in credit activities. Consistent with the research results of Mahfuzah Salim and Dr Raj Yadav (2012), the author also shows that the size of the company has a positive impact on the performance of the company. The larger the company size, the higher the corporate profits. *Hypothesis H5: Firm size (SIZE) has a positive effect on firm's profit margin.*

**Inflation Rate – INF:** According to Comley (2015), inflation occurs when the purchasing power of money decreases due to an increase in the price level of goods and services in the economy. The formula for calculating the inflation rate according to the consumer price index (CPI) is as follows:



$$\text{Inflation rate in period } t = \frac{\text{Consumer Price Index}_n - \text{Consumer Price Index}_{n-1}}{\text{Consumer Price Index}_{n-1}} \times 100\%$$

In there n: n-th year. According to Tran Viet Dung and Bui Dan Thanh (2021), the inflation rate is a macro factor that affects the capital structure of enterprises. When the inflation rate of the economy is high, the Government will require tightening credit to curb inflation, causing the lending interest rate of banks to increase. This leads to a shortage of capital for construction enterprises, adversely affecting the profits of enterprises. However, in another aspect, inflation creates cheap capital, contributing to economic development. A typical example is that in the period 1984–1997, China accepted an average inflation of 10.98% to raise the amount of capital from issuing money to 3235.71 billion yuan. This helps China's GDP grow 3.23 times (According to financial statistics of the IMF), becoming the world's 2nd economic power after the United States (According to the ranking of the list of countries by gross domestic product of the International Monetary Fund in 2021). If Vietnam's economy maintains inflation at a moderate level, construction businesses can also develop and increase profits through the stability of the country's macro-economy. **Hypothesis H6: Inflation rate (INF) has a positive effect on firm's profit rate** (Table 1).

**Table 1.** Description of the model's variables, measurement methods and hypotheses

Variable	Description	Measurement	Hypotheses
<b>Dependent variable</b>			
<b>ROA</b>	Return on Assets	$ROA = \frac{\text{Earning After Tax}}{\text{Total Assets}}$	
<b>Biến độc lập</b>			
<b>SDA</b>	Short – term debt to total assets	$SDA = \frac{\text{Current Liabilities}}{\text{Total Assets}}$	–
<b>LDA</b>	Long – term debt to total assets	$LDA = \frac{\text{Long term Debt}}{\text{Total Assets}}$	–
<b>LQ</b>	Liquidity	$LQ = \frac{\text{Current Assets}}{\text{Current Liabilities}}$	–
<b>GROWTH</b>	Revenue growth rate	$GROWTH = \frac{\text{Net Revenue}_n - \text{Net Revenue}_{n-1}}{\text{Net Revenue}_{n-1}} \times 100\%$	+
<b>SIZE</b>	Firm size	$SIZE = \text{Log}_e(\text{Total Assets})$	+
<b>INF</b>	Inflation Rate	$\text{Inflation rate in period } t = \frac{\text{Consumer Price Index}_n - \text{Consumer Price Index}_{n-1}}{\text{Consumer Price Index}_{n-1}} \times 100\%$	+

Source: Compiled by the author

### 3.4 Model Estimation Method

To evaluate the impact of foreign ownership on liquidity risk, the authors will make model estimation according to Bayesian approach. To conduct a Bayesian analysis, a priori information is required for the research model, but since most of the prior research was performed using a frequency approach, a priori information is not available. However, the research data of 792 observations is quite large, so the a priori information does not have a great influence on the posterior distribution. In this case, Block et al. (2011) proposed a standard Gaussian distribution with different a priori information (simulation of a priori information) and carried out Bayesian factor analysis to choose a simulation with the best priori news.

**Table 2.** Simulation of a priori information

Rational function	ROA $\sim N(\mu, \sigma)$
A priori distribution	
Simulation 1	$\alpha \sim N(0, 1)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 2	$\alpha \sim N(0, 10)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 3	$\alpha \sim N(0, 100)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 4	$\alpha \sim N(0, 1000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 5	$\alpha \sim N(0, 10000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$

Source: Compiled by the author

The simulations in Table 2 show decreasing levels of a priori information with Simulation 1 having the strongest a priori information and Simulation 5 having the weakest a priori information.

In the next step, the research team carried out Bayesian regression for the above simulations, then performed Bayesian factor analysis (Bayes Factors) and Bayes test model (bayestest model). These are the techniques proposed by StataCorp LLC (2019) to select the simulation with the best a priori information. Basically, the Bayesian factor will provide a tool to compare the probability of a particular hypothesis (a priori information) to the probability of another hypothesis. It can be understood as a measure of the strength of evidence in favor of a theory among competing (information a priori) theories. Accordingly, Bayesian analysis will provide average Log BF (Bayes Factor), Log ML (Marginal Likelihood) and average DIC (Deviance Information Criterion - information bias); The posterior Bayesian test will help compare the posterior probability of the simulations with different a priori information, accordingly, based on the research data

combined with the proposed a priori information, we will choose the simulation has the greatest posterior probability  $P(M|y)$ .

In summary, in this study, the research team will build 5 simulations with 5 different a priori information, and Bayesian factor analysis and posterior Bayesian test will help to choose a simulation with suitable a priori information. The simulation selected will be the one with the largest Log BF, Log ML average, minimum DIC mean and the largest  $P(M|y)$ .

## 4 Research Results and Discussion

### 4.1 Results

The simulation selected will be the one with the largest Log BF, Log ML average, minimum DIC mean, and the largest  $P(M|y)$ . Simulation 2 has the largest Log BF, however, Avg DIC is not as good as 3, 4, 5, in addition, Avg Log (ML) of simulation 2 does not show its superiority compared to other simulations, so we continue to analyze the Bayes test model. The results show that simulation 2 has a higher posterior probability than the other simulations, so simulation 2 with a priori distribution  $N(0,10)$  will be selected (Table 3).

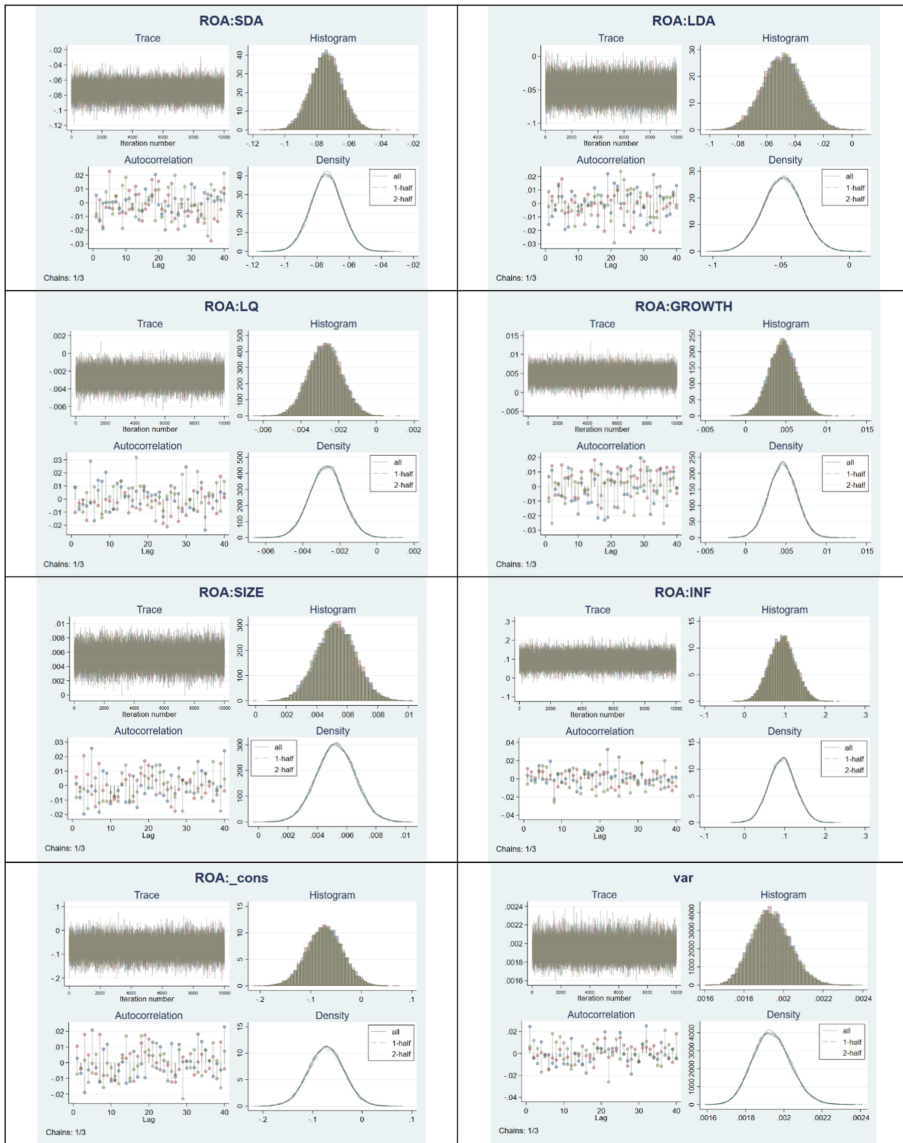
**Table 3.** Bayes Factor analysis results

	Chains	Avg DIC	Avg log (ML)	Log (BF)	$P(M y)$
Simulation1	3	4610.8	-2.32E+03		0.000
Simulation2	3	4593.3	-2.32E+03	8.129	0.893
Simulation3	3	4591.0	-2.32E+03	6.004	0.107
Simulation4	3	4591.4	-2.33E+03	-0.960	0.000
Simulation5	3	4591.4	-2.33E+03	-9.039	0.000

*Source: Calculations of the author*

Bayesian analysis is simulated through a Markov chain Monte Carlo (MCMC), so to ensure the stability of Bayesian regression, the MCMC series must converge, which means that the MCMC series must ensure stationarity. StataCorp LLC (2019) proposes that the MCMC series convergence test can be conducted through the convergence diagnostic graph.

According to StataCorp LLC (2019), the MCMC series convergence diagnostic graph includes trace plot, histogram, autocorrelation, and density estimation. The trace plot helps to track the historical display of a parameter value over the iterations of the series, Fig. 1 shows the trace plot fluctuates around the mean value, so the MCMC series is stationary, that is, reaching convergence conditions. Besides, the autocorrelation chart in the graphs only fluctuates around the level below 0.02, according to StataCorp LLC (2019) the autocorrelation chart fluctuates around the level below 0.02, showing the agreement with the density simulate the distribution and reflect all delays that are within



**Fig. 1.** Convergence diagnostic graph Source: Calculations of the author

the effective limit. According to StataCorp LLC (2019), the posterior distribution plot and density estimate show that the simulation of the shape of the normal distribution of the parameters, the histogram shape is uniform, it can be concluded that Bayesian regression ensure stability. Thus, the results from Fig. 1 show that the MCMC series meets the convergence condition.

In addition to graphical convergence diagnostics, StataCorp LLC (2019) also recommends testing through Mean Acceptance Rate; Average minimum efficiency; and

**Table 4.** Regression results

	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
SDA	-0.074	0.010	0.000	-0.074	-0.093	-0.055
LDA	-0.049	0.014	0.000	-0.049	-0.076	-0.021
LQ	-0.003	0.001	0.000	-0.003	-0.004	-0.001
GROWTH	0.005	0.002	0.000	0.005	0.001	0.008
SIZE	0.005	0.001	0.000	0.005	0.003	0.008
INF	0.094	0.033	0.000	0.094	0.028	0.159
_cons	-0.073	0.035	0.000	-0.073	-0.141	-0.004
var	0.002	0.000	5.70E+07	0.0019	0.0018	0.0021
Avg acceptance rate	1.000					
Avg efficiency min	0.984					
Max Gelman-Rubin Rc	1.000					

Source: Calculations of the author

Gelman-Rubin Rc max. Table 4 shows that the model's acceptance rate reaches 1, the model's minimum efficiency is 0.984, far exceeding the allowable level of 0.01. In addition, the maximum Rc value of the coefficients is 1, Gelman and Rubin (1992) argue that the diagnostic value Rc of any coefficient of the model greater than 1.2 will be considered non-convergent. Thus, the values in Table 4 show that the MCMC series of the model satisfy the convergence requirements.

The regression results in Table 4 have determined that the variables SDA, LDA, LQ have a negative impact on the profit margin of construction enterprises while the variables GROWTH, SIZE, INF have a positive effect on the rate of return. Besides determining the sign of the regression coefficients, unlike the frequency method, the Bayesian approach also allows us to calculate the probability of these effects.

## 4.2 Discussion

**Short-Term Debt to Total Assets – SDA:** Research results show that SDA has a negative relationship with ROA, in line with the author's initial expectation. The higher the short-term debt ratio, the lower the profit margin of the construction industry. This result is consistent with the experimental studies of Doan Ngoc Phuc (2018), Pouraghajan and Malekian (2012), Nguyen Thi Dieu Chi (2018). When construction enterprises prioritize using short-term financing in the course of business operations, it will create short-term payment pressure, negatively affecting the profits of enterprises. Moreover, when the ratio of short-term debt to total assets is high, it means that the enterprise maintains a relatively high amount of short-term assets, resulting in the enterprise not effectively using short-term assets.

**Long-Term Debt to Total Assets – LDA:** Research results show that LDA has a negative relationship with ROA, in line with the author's initial expectation. Construction enterprises often have relatively large capital requirements and long business cycles. Long-term debt is considered an essential source of capital in the capital structure of construction enterprises. However, overusing this capital will cause businesses to face high interest rates when the market economy fluctuates. High interest rates will lead to an increase in the burden of interest expenses or interest on borrowed capital, negatively impacting business performance. This result is consistent with the study of Pouraghajan and Malekian (2012), Mesquita and Lara (2003), Abor (2005).

**Liquidity – LQ:** Research results show that LQ has a negative relationship with ROA, in line with the author's initial expectation. However, this result is contrary to the study of the authors Dang Phuong Mai (2016), Nguyen Thi Dieu Chi (2018). When the expenses of maintaining liquidity are unstable, leading to costs incurred in the business's operation and not ensuring the ability to pay short-term debts of the enterprise. The characteristics of enterprises in the construction industry are slow-moving short-term assets and large receivables. When current assets are larger than current liabilities, it does not guarantee that current assets will be able to pay short-term liabilities when they come due. Current assets are highly liquid assets such as cash, so when short-term assets increase, the profit margin of the business also decreases.

**Revenue Growth Rate – GROWTH:** Research results show that GROWTH has a positive relationship with ROA, in line with the author's initial expectation. For businesses, the business's operational efforts are often reflected in revenue growth. Enterprises with higher revenue growth contribute to increased profits and increase business value. This result is consistent with the study of Pouraghajan and Malekian (2012), Tian and Zeitun (2007).

**Firm Size - SIZE:** Research results show that SIZE has a positive relationship with ROA, in line with the author's initial expectation. As the business grows in size through increasing assets, the value of the business will also increase. This result is similar to the study of Gleason (2000), Abor (2005) and Ahmad (2012). Large enterprises in the construction industry often have high competitiveness, so it is easier for businesses to carry out activities to expand their scale. This increases the profit margin and the value of the business.

**Inflation Rate – INF:** Research results show that INF has a positive relationship with ROA, in line with the author's initial expectation. This result is consistent with the study of Wanjohi (2003), but contradicts with the study of Vena (2012). If Vietnam's economy maintains moderate inflation, construction enterprises will have to take adjustment measures to adapt to changes in the economy. Since then, businesses still develop and increase profits through the stability of the macro economy.

## 5 Conclusion and Policy Implications

### 5.1 Conclusion

In the period of 2010–2020, construction enterprises in Vietnam increased their profits mainly thanks to the increase in asset size, business development, and cost reduction. The inflation rate has a positive impact on the profit rate of enterprises because the Government has tried to recover the economy after the global economic crisis in 2008 with various measures to control inflation, leading to GDP growth returned, leading to an increase in inflation.

### 5.2 Policy Implications

**Short-Term Debt:** Enterprises should prepare well loan documents, transparent financial statements, etc. to increase access to capital and reduce transaction costs from banks or credit institutions. In addition, businesses need to make sure to make payments to suppliers in accordance with the agreement, the purpose of which is to increase the reputation and position of the business.

**Long-Term Debt:** One of the solutions to provide effective long-term capital today is the form of financial leasing because this is a highly safe and effective form of financing for the transaction parties. To implement this project, the enterprise must have a viable finance lease project, a healthy financial situation and financial ability to participate in the lease project. Especially, enterprises need to have an effective production and business plan. For construction companies listed on the HSX, bond issuance is one of the long-term capital mobilization operations. Therefore, in order to successfully issue bonds, enterprises need to prove their operational capacity to have high profits, ensure debt repayment and interest expenses for investors.

**Liquidity:** When it is unable to pay its short-term debts, a business can switch from short-term debt to long-term debt, which results in smaller monthly payments and gives the business more time to pay off. In addition, businesses can keep a relative level of liquidity, not too low by controlling overhead costs or selling unnecessary assets to ensure liquidity.

**Revenue Growth Rate:** Enterprises should focus on product quality to improve their reputation and increase their ability to meet the needs and desires of customers. In addition, expanding the business into new markets or areas and scaling up also helps in a high rate of revenue growth.

**Enterprise Size:** Enterprises can increase their business size through increasing capital to expand their business scale, investing in fixed assets to build more branches in big cities and densely populated area to increase the popularity of the business among new customers.

## References

Abor, J.: The effect of capital structure on profitability: an empirical analysis of listed firms in Ghana. *J. Risk Finance* 6(5), 438–445 (2005)

- Ahmad, Z., Abdullah, N.M.H., Roslan, S.: Capital structure effect on firms performance: focusing on consumers and industrials sectors on Malaysian firms. *Int. Rev. Bus. Res. Pap.* **8**(5), 137–155 (2012)
- Thanh, B.D.: Cấu trúc vốn và vốn luân chuyển tác động đến hiệu quả quản trị tài chính của các doanh nghiệp nhỏ và vừa trên địa bàn thành phố Hồ Chí Minh. Luận án Tiến Sĩ Kinh tế, Trường Đại học Ngân hàng thành phố Hồ Chí Minh (2016)
- Comley, P.: *Inflation Matters: Inflationary Wave Theory. Its Impact on Inflation Past and Present and the Deflation Yet to Come*, Pete Comley, London (2015)
- Mai, D.P.: Giải pháp tài cấu trúc tài chính các doanh nghiệp trong ngành thép ở Việt Nam. Luận án Tiến Sĩ Kinh tế, Học viện Tài chính (2016)
- De Mesquita, J.M.C., Lara, J. E.: Capital structure and profitability: the Brazilian case. In: *Academy of Business and Administrative Science Conference*, Vancouver, Canada, pp. 11–13, July 2003
- Phúc, D.N.: Ảnh hưởng của cấu trúc vốn đến hiệu quả hoạt động kinh doanh của doanh nghiệp sau cổ phần hóa ở Việt Nam. *Tạp chí nghiên cứu kinh tế, Viện Kinh tế Việt Nam* **1**(476), 11–16 (2018)
- Donaldson, G.: *Corporate Debt Capacity: A Study of Corporate Debt Policy and the Determination of Corporate Debt Capacity*. Division of Research, Graduate School of Business Administration, Harvard University, Boston (1961)
- Gill, A.: The effect of capital structure on profitability: evidence from the United States. *Int. J. Manag.* **28**(4) (2011)
- Githaiga, P.N., Kabiru, C.G.: Debt financing and financial performance of small and medium size enterprises: evidence from Kenya. *J. Econ. Finance Account.* **2**(3), 473–481 (2015)
- Gleason, K.C., Mathur, L.K., Mathur, I.: The interrelationship between culture, capital structure, and performance: evidence from European retailers. *J. Bus. Res.* **50**(2), 185–191 (2000)
- Goddard, J., Molyneux, P., Wilson, J.: The profitability of European banks: a cross sectional and dynamic panel analysis. *Manch. Sch.* **72**(3), 363–381 (2004)
- Kraus, A., Litzenberger, R.H.: A state-preference model of optimal financial leverage. *J. Financ.* **28**(4), 911–922 (1973)
- Modigliani, F., Miller, M.H.: The cost of capital, corporation finance and the theory of investment. *Am. Econ. Rev.* **48**(3), 261–297 (1958)
- Modigliani, F., Miller, M.H.: Corporate income taxes and the cost of capital: a correction. *Am. Econ. Rev.* **53**(3), 433–443 (1963)
- Molyneux, P., Thornton, J.: Determinants of European bank profitability: a note. *J. Bank. Finance* **16**(6), 1173–1178 (1992)
- Myers, S.C.: Determinants of corporate borrowing. *J. Financ. Econ.* **5**(2), 147–175 (1977)
- Myers, S.C.: The capital structure puzzle. *J. Finance* **39**(3), 575–592 (1984)
- Nguyễn, T.D.C.: Tác động của cấu trúc vốn nợ tới hiệu quả tài chính: Nghiên cứu điển hình các doanh nghiệp dịch vụ Việt Nam. *Tạp chí Khoa học Công nghệ* **60**(11) (2018)
- Nunes, P.J.M., Serrasqueiro, Z.M., Sequeira, T.N.: Profitability in Portuguese service industries: a panel data approach. *Serv. Ind. J.* **29**(5), 693–707 (2009)
- Pouraghajan, A., Malekian, E.: The relationship between capital structure and firm performance evaluation measures: evidence from the Tehran stock exchange. *Int. J. Bus. Commer.* **1**(9), 166–181 (2012)
- Salim, M., Yadav, R.: Capital structure and firm performance: evidence from Malaysian listed companies. *Proc. Soc. Behav. Sci.* **65**, 156–166 (2012)
- Dũng, T.V., Thanh, B.D.: Các nhân tố ảnh hưởng đến cấu trúc vốn của các doanh nghiệp niêm yết trên Thị trường chứng khoán Việt Nam. *Tạp chí Khoa học và Đào tạo Ngân hàng, Học viện Ngân hàng* (2021)
- Vena, H.: *The Effect of Inflation on the Stock Market Returns of the Nairobi Securities Exchange*. Unpublished master's thesis, The University of Nairobi (2012)



- Wanjohi, J.C.: Determinants of Commercial Banks Profitability in Kenya: The Case of Kenya Quoted Banks. Unpublished master's thesis, The University of Nairobi (2003)
- Zeitun, R., Tian, G., Keen, S.: Macroeconomic determinants of corporate performance and failure: evidence from an emerging market the case of Jordan. *Australas. Account. Bus. Finance J.* **1**(4), 44–61 (2007)



# Determinants of Small and Medium Enterprises' Capital Intensity: The Case in Vietnam

Nhan Truong Thanh Dang, Van Dung Ha, and Van Tung Nguyen<sup>(✉)</sup>

Banking University Hochiminh City, Ho Chi Minh City, Vietnam  
{nhandtt, dunghv, tungnv}@buh.edu.vn

**Abstract.** Generally, there is still limited research, if any, which aims to identify and analyze determinants of the capital intensity of firms in Vietnam, including SMEs. Considering this research gap, this paper focuses on identifying various determinants of SMEs' capital intensity across different industries in Vietnam. The influence level of different determinants, such as firm size, firm age, managers' gender, and education, will also be evaluated using the Bayesian statistics method. The result findings show that: bigger firms in terms of the total labor force have lower capital intensity ratios; firms with larger liability have higher capital intensity ratios; older firms have higher capital intensity ratios; SMEs with male managers have higher capital intensity ratios. Additionally, firms with university-educated managers and vocational trained managers will have a lower capital intensity ratio than those with not formally educated managers. It is highly recommended that SMEs' managers invest more in standardizing operation processes, quality management systems, and training programs along with expanding business activities. Additionally, firms should continually innovate their working process and apply advanced technology to adapt to rapid environmental changes and compete with their rivals.

**Keywords:** capital intensity · determinants · employment · liability · SMEs

**JEL Classification:** G21 · C11 · C58

## 1 Introduction

Despite a large amount of previous research related to the topic of capital intensity, only a few studies focused on the determinants of capital intensity degree (Shojaie and Tehranchian 2018). Capital intensity, which could be calculated by total assets divided by sales, is viewed as an important factor following economic growth theories and international economics. Therefore, acknowledging factors that effectively enhance capital intensity can help foster the growth of firms, including small and medium enterprises (SMEs), as well as help governments develop better policies to promote development and international trade.

Especially based on the perception that the ultimate objective of financial managers and board of directors is to maximize firm value (Andrew et al. 2007), an investigation of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 264–273, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_18](https://doi.org/10.1007/978-3-031-35763-3_18)

the impact of capital intensity on firm value as well as the determinants of capital intensity should enlighten researchers and practitioners across different industries. However, the literature concerning capital intensity still provides inconclusive and mixed results and requires more empirical evidence across different contexts around the globe (Lee 2010; Judzik and Sala 2015; Shojaie and Tehranchian 2018).

Regarding the Vietnamese context, in recent years, capital intensity has been increasingly discussed following different research objectives, such as analyzing the influence of determinants of labor productivity varying among different firm sizes and different levels of capital intensity (Vu 2022) and investigating the interrelationship among firms' productivity, capital intensity and the decision-making of the adoption (Ni 2015). Generally, there is still limited research, if any, which aims to identify and analyze determinants of the capital intensity of firms in Vietnam, including SMEs. Considering this research gap, this paper focuses on identifying various determinants of SMEs' capital intensity across different industries in Vietnam. The influence level of different determinants, such as firm size, firm age, managers' gender, and education, will also be evaluated using the Bayesian statistics method. Consequently, managerial implications would be provided to support firms to improve the level of capital intensity and foster economic growth accordingly.

This research paper can provide a theoretical contribution to the literature development about the emerging topic of enterprises' capital intensity, with a concentration on SMEs in a developing country such as Vietnam. This empirical study can be considered as a reply to the call raised by other academies for more research on causing factors of capital intensity (Lee 2010; Judzik and Sala 2015; Shojaie and Tehranchian 2018) as well as a notable reference for future research focusing on capital intensity in emerging countries. Practically, this study can help to provide managerial implications for the purpose of enhancing firms' capital intensity and eventually contributing to economic growth.

## 2 Literature Review

### Capital Intensity Definitions

There have been different perspectives about capital intensity definitions.

Lubatkin and Chatterjee (1994) define capital intensity as representative of a firm's operating leverage. Levels of capital intensity are different among different industries. Examples of capital-intensive industries are mining, utilities, airlines, railroads, cruise lines, hotels, and restaurants (Lee 2010).

According to Shojaie and Tehranchian (2018), capital intensity or intensity of production factors utilization is defined as the relative measures of two production factors utilized for producing a good to another good. The capital intensity or capital-labor ratio ( $k/l$ ) demonstrates the labor force and capital used in production (Shojaie and Tehranchian 2018).

Capital-labor ratio indicates the amount of capital needed for creating new jobs in industry; therefore, it can show capital-intensive or labor-intensive technology.

As mentioned by Lee (2010), capital intensity is an economic term indicating how much capital is used for production compared to two other factors, especially labor.

From the financial perspective, capital intensity is considered an indicator that shows that the industry uses a large amount of capital for production.

Another definition is introduced by Muzakki and Darsono (2015), which states that capital intensity represents the amount of capital invested by a firm in the form of fixed assets. More investment in the form of fixed assets is one of the firm's strategies for implementing tax avoidance practices. Almost all fixed assets lead to depreciation, meaning that depreciation charges will arise, resulting in more financial burdens for the firm. The large deduction from profit in tax calculations would help lower the pre-tax profit and, thereby, reduce the tax the company must pay. Therefore, the higher value of capital intensity in a firm, the higher the probability that the firm will practice tax avoidance.

In addition to operating leverage (Lubatkin and Chatterjee 1994), a company's capital intensity can be evaluated by calculating how many assets are required to produce a dollar of sales, which is total assets divided by sales (Frankenfield 2020). This is the inverse of the asset turnover ratio, an indicator of the efficiency with which a firm uses its assets to generate revenue. Such a definition is similar to the explanation of Finance Management (2021), which states that capital intensity indicates the efficacy with which the capital and assets of the firm are utilized for production.

This research applied the definition of capital intensity as evaluated by calculating how many assets are required to produce a dollar of sales, which is total assets divided by sales. This definition indicates the efficiency with which the capital and assets of the firm are employed for production.

### **Theoretical Background Related to Capital Intensity**

Economic growth theories can be used to explain the observed realities of growth on a global scale (Shojaie and Tehranchian 2018). Adam Smith was viewed as the first theorist who suggested the official description of economic growth in his book named "Wealth of Nations." His research viewed division of labor, saving, capital concentration, technology improvement, expertise and market development as effective factors in growth and development. He believed capital concentration is necessary for economic development, which can be achieved through a saving increase. Economic growth can be restricted by the scarcity of resources (Shojaie and Tehranchian 2018).

Different growth models were developed based on the economic perspective that technical progress can be obtained through investment in research and development and the creation of ideas (Romer 1990; Grossman and Helpmen 1991; Aghion and Howitt 1992).

Harrod – Domar growth model concentrated on the role of investment, in terms of assets especially. According to their model, investment and capital accumulation play the main roles in economic growth and development (Shojaie and Tehranchian 2018). Investment has two effects: Firstly, the generation of income/sales, and second, enhancing the production capacities.

Despite some criticisms about all the Harrod – Domar growth model, such as: Impossibility of substitution of labor and capital, the hypotheses of inflexibility, close economy and constant technology, the role of capital accumulation in fostering economic

growth through saving has been highlighted in later development theories (Shojaie and Tehranchian 2018).

### **Previous Studies Related to Determinants of Capital Intensity**

Hurdle (1974) investigated the relation between leverage, market structure, risk and profitability. Based on developing a theoretical framework for these variables and examining the model for 2228 production firms in the United States during the 1960s, the study included capital intensity in the risk equation. It concluded that high capital intensity is correlated with low risk.

In the study of Lubatkin and Chatterjee (1994), the relation between diversity strategy and risk in 246 stock market firms during 1970–1984 was explored. Based on the research findings, there is a negative relation between capital intensity and risk.

In the research of Judzik and Sala (2015) about the determinants of capital intensity in Japan and the United States, the autoregressive distributed lag (ARDL) method was applied for time series data during 1970–2011. Independent variables included relative costs of the production function, participation rate of production factors, trade openness degree and direct tax of household and trade taxes. Results indicate that in the United States, demand-side factors such as participation rate of capacity have more impact on capital intensity.

Shojaie and Tehranchian (2018) examined the determinants of capital intensity in Iran and China. The research found that in Iran, the openness degree has the largest effect on capital intensity in the short run. Nonetheless, in the long run, relative costs of production factors have the most significant effect. Meanwhile, in China, in both the short and the long run, the participation rate of production factors has the largest impact on capital intensity.

[https://www.dropbox.com/s/xlrdryqg9uz54bc/root%2C%2BJournal%2Bmanager%2C%2B13\\_IJEFI\\_6187%2BTehranchian%2Bokey\\_20180215\\_V1.pdf?dl=0](https://www.dropbox.com/s/xlrdryqg9uz54bc/root%2C%2BJournal%2Bmanager%2C%2B13_IJEFI_6187%2BTehranchian%2Bokey_20180215_V1.pdf?dl=0).

Regarding the relation between firm age and capital intensity, Boring (2020) concluded that the average capital intensity is highest among the youngest firms and lowest among the oldest firms. The finding can explain that the proportion of highly skilled workers is highest among the youngest firms and lowest among the oldest firms. At the same time, there is a positive correlation between this proportion and the average capital intensity. The general conclusion from this study is that newly founded firms are more capital-intensive than incumbent firms, on average.

In the report prepared by Måns Söderbom, UNIDO Research Fellow at the CSAE, in co-authorship with Francis Teal (2001), substantial differences in capital intensity over the size range of firms were discussed. The report highlighted that capital costs differ by the size of the firm, given the influence of size on access to capital. The report concluded a large dispersion of capital intensity by firm size in the data related to manufacturing firms in Ghana. According to the report's explanation, this size dispersion was not due to technology, as it does not reflect the use of more capital-intensive technology in some sectors and is not an outcome of changes in labor productivity during the surveys.

In terms of the impact of managers' education on capital intensity, which is strongly related to firm growth, Queiro (2016) found that firms that switch to more educated managers experience a sharp increase in growth relative to comparable firms. More educated managers are also more likely to use incentive pay and are more likely to develop new

products and services and incorporate new technologies. The findings suggest that the impact occurs through technology adoption and human resource management.

Managers' gender is also a considerable variable that can affect firms' financial decisions, associated with firms' capital intensity. Some of the most widely identified determination factors in the literature regarding gender differences in financial behavior are risk attitudes (Croson and Gneezy 2009; Eckel and Grossman 2008), financial knowledge (Fonseca et al. 2012; Lusardi and Mitchell 2008, 2011), overconfidence (Barber and Odean 2001; Hira and Mugenda 2000) and cognitive style (Epstein 2003; Sladek et al. 2010). Several previous studies found that men have more confidence in their financial skills than females (Chen and Volpe 2002; Fonseca et al. 2012; Lusardi and Mitchell 2008, 2011; Zissimopoulos, Karney and Rauer 2008). Some studies focusing on gender-related dissimilarities in investment patterns suggested that female investors were likely to have less confidence in their investment decisions than male investors under similar circumstances (Agnew et al. 2003; Barber and Odean 2001; Webster and Ellis 1996).

### 3 Methodology

A dataset of more than 2,500 small and medium enterprises in Vietnam is used to identify the determinants of SMEs capital intensity. The dataset was compiled by the Central Institute for Economic Management (CIEM), the Institute of Labor Science and Social Affairs (ILSSA), the United Nations University's Institute for World Development Economics (UNU-WIDER) and the Faculty of Economics (DOE) of the University of Copenhagen from 2007 to 2015.

Based on the literature review, the article introduces the following variables into the model:

$$CIR_{it} = \beta_1 + \beta_2 \text{ Employment}_{it} + \beta_3 \text{ Liability}_{it} + \beta_4 \text{ Age}_{it} + \beta_5 \text{ Gender}_{it} + \beta_6 z1_i + \beta_7 z2_i + \beta_8 z3_i + \varepsilon_i$$

#### Where:

#### Dependent Variable

CIR is the capital intensity ratio, which is measured by the ratio of total Assets to Sales of firms.

#### Explanatory Variables

Employment is total full-time employment, measured in natural logarithm.

Liability is total firm liability, measured in natural logarithm.

Age is the number of years of establishment as a proxy of firm age.

Gender is the manager's gender, and gender equals 1 if firm's manager is male and zero otherwise.

$z1$ ,  $z2$ , and  $z3$  are dummy variables for manager educational level.  $z1$  receives 1 if the manager graduated from university, and otherwise 0.  $z2$  equals 1 if the manager graduated from vocational training, and otherwise 0.  $z3$  equals the value of 1 if the manager got training without degree, and otherwise 0.

For precise estimation results, Bayesian inference is employed. With prior information and data, Bayesian estimation gives more reliable estimates than other methods in the case of small samples, lacking data (Baldwin and Fellingham 2013).

## 4 Findings and Discussions

Bayesian coefficients and Bayesian estimation models will be selected by the maximum mean for Log BF, Log (ML), P(M/y), and the minimum of DIC (Penny et al. 2007).

**Table 1.** Bayesian factor test and model test

	Chan	lnLP			
		Avg DIC	Avg log (ML)	Avg log (BF)	P(M/y)
Simulation 1	3	1.71e+04	-8.61e+03		0.0000
Simulation 2	3	1.71e+04	-8.61e+03	<b>13.9783</b>	<b>0.9989</b>
Simulation 3	3	1.71e+04	-8.61e+03	7.1268	0.0011
Simulation 4	3	1.71e+04	-8.61e+03	-1.7954	0.0000
Simulation 3	3	1.71e+04	-8.61e+03	-11.0174	0.0000

Source: Author's calculation

Based on the results of Table 1, Model 1 is selected.

The research paper will analyze the sensitivity through five simulations of normally distributed a priori information to select the appropriate a priori information for a large sample size as follows (Tables 2 and 3):

**Table 2.** Likelihood model

	lnLP $\sim N(\mu, \delta)$
Prior distributions:	
Simulation 1	$\alpha_i \sim N(0, 1)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 2	$\alpha_i \sim N(0, 10)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha_i \sim N(0, 100)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 4	$\alpha_i \sim N(0, 1000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha_i \sim N(0, 10000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
$i = 1, 2, 3, 4, 5$	

**Table 3.** Bayesian simulation results

CIR						
	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
Employment	-.1081394	.0196393	.000113	-.1080358	-.1466882	-.0699096
Liability	.0929454	.0101534	.000059	.0929533	.0731652	.1129489
Age	.0109433	.0017839	.00001	.0109554	.007435	.0144449
Gender	.0339139	.0344699	.000201	.033812	-.0331482	.1013312
z1	-.0599472	.0530153	.000307	-.0597453	-.164087	.0447204
z2	-.0740107	.0484152	.00028	-.0739422	-.1693282	.0203036
z3	.0783423	.0471359	.000274	.0784403	-.0144061	.1699928
_cons	-7.193515	.0608311	.000354	-7.193582	-7.312277	-7.074095
var	1.429789	.0275693	.00016	1.429464	1.37617	1.484958

Number of obs = 10,000  
 Avg acceptance rate = 1  
 Avg efficiency: min = .9838  
 Max Gelman-Rubin Rc = 1

Source: Author's calculation

After the model is estimated, the MCMC convergence needs to be considered. According to Gelman and Rubin (1992), Brooks and Gelman (1998), diagnostic Rc values greater than 1.2 for any model parameter are considered non-convergent. In practice,  $Rc < 1.1$  is often used to declare convergence. Therefore, the Max Gelman-Rubin Rc value is smaller than 1.1, indicating that the MCMC convergence is acceptable for Bayesian analysis.

The paper also checks the robustness of the results by testing the sensitivity of the posterior estimates. The results of the certainty test show that the posterior estimates are not significantly different in terms of the posterior mean, MCSE and confidence intervals when the normal base values for all parameters are adjusted from  $-0.5$  to  $0.5$  with  $0.1$  interval.

As per the results, it is found that bigger firms, in terms of the total labor force, have a lower capital intensity ratio. This finding is supported by the research of Måns Söderbom, UNIDO Research Fellow at the CSAE, in co-authorship with Francis Teal (2001), which concluded substantial differences in capital intensity over the size range of firms. The firm size can be indicated by total labor force or total equity, which are associated with the efficiency with which a firm can use its assets to generate revenue. Bigger firms tend to invest more in ensuring standardization of working processes, quality management, and training activities due to their expanded business scopes. Such efforts are associated with their efficiency in asset utilization to generate revenues, indicating a lower level of capital intensity. This result can also be explained by Harrod – the Domar growth model, which emphasized the role of investment, regarding assets especially, in economic growth and development via the generation of income/sales and enhancing the production capacities (Shojaie and Tehranchian 2018).



The results also show that firms with larger liability have a higher capital intensity ratio. This indicates that the debt level also has a correlation in the same direction as capital intensity. The debt level is normally related to the degree of risk a firm is willing to take. Risk-averse mindsets usually make business leaders more careful in managing their liability or reducing liability/debt levels. This research's result is not similar to the findings of Hurdle (1974) and Lubatkin and Chatterjee (1994), which proved a negative relation between capital intensity and risk. According to this research's findings, an enterprise's high level of liabilities may be partially due to its broad investment in assets, especially fixed assets, which simultaneously increases capital intensity.

Another remarkable finding is that older firms have a higher capital intensity ratio. This is dissimilar from the conclusion of Boring (2020) that the average capital intensity is highest among the youngest firms and lowest among the oldest firms. According to previous studies, on average, newly founded firms are more capital-intensive than incumbent firms. However, in this research context, it appears that older firms are not as efficient as new firms regarding their ability to utilize assets to produce sales. Older firms may encounter difficulties in innovating their traditional working process to adapt to rapid environmental changes and compete with young and creative firms.

The research findings also reveal that SMEs with male managers have a higher capital intensity ratio. Previous studies discussed gender differences in terms of risk attitudes (Croson and Gneezy 2009; Eckel and Grossman 2008), financial knowledge (Fonseca et al. 2012; Lusardi and Mitchell 2008, 2011), overconfidence (Barber and Odean 2001; Hira and Mugenda 2000) and cognitive style (Epstein 2003; Sladek et al. 2010). Several previous studies concluded that men have more confidence in their financial skills than females (Chen and Volpe 2002; Fonseca et al. 2012; Lusardi and Mitchell 2008, 2011; Zissimopoulos, Karney and Rauer 2008). The male managers' confidence in their financial skills and knowledge may lead to their high intention to invest extensively in fixed assets, possibly leading to a higher capital intensity ratio.

Last but not least, the findings show that firms that have managers with university graduation and vocational training will have a lower capital intensity ratio compared to those firms with no-formal educational managers. Besides, firms with managers with training without degrees will have a higher capital intensity ratio than those with no-formal educational managers. This finding is in line with the argument of Queiro (2016) that firms that switch to more educated managers experience a sharp increase in growth via developing new products and services and incorporating new technologies. Incorporating new technologies and innovation directed by educated managers with degrees and qualifications would possibly enhance the efficiency level in utilizing assets and producing revenues, indicating a lower capital intensity ratio.

## 5 Conclusion

To recapitulate, based on the application of the Bayesian estimation method, this research analyzed the dataset of more than 2,500 SMEs in Vietnam to identify determinants of capital intensity in these firms. Bayesian coefficients and Bayesian estimation models were selected by the maximum mean for Log BF, Log (ML), P(M/y), and the minimum of DIC. The Max Gelman-Rubin  $R_c$  value was smaller than 1.1, indicating that the MCMC convergence is acceptable for Bayesian analysis.

The result findings show that: bigger firms in terms of the total labor force have lower capital intensity ratio; firms with larger liability have higher capital intensity ratios; older firms have higher capital intensity ratios; SMEs with male managers have higher capital intensity ratios. Additionally, firms with university graduation and vocational training managers will have a lower capital intensity ratio compared to those with no-formal educational managers. In comparison, firms with managers with training without degrees will have a higher capital intensity ratio than those with no-formal educational managers.

This research can be seen as a contribution to the literature development about the topic of enterprises' capital intensity, with the examination of SMEs in developing countries such as Vietnam. This study can also be viewed as a response to the call raised by previous studies for more research on determinants of capital intensity (Lee 2010; Judzik and Sala 2015; Shojaie and Tehranchian 2018), as well as a notable reference for future related research.

Based on the findings, there are some managerial implications that business leaders can consider. Firstly, it is highly recommended that SMEs' managers invest more in standardizing operation processes, quality management systems and training programs, along with expanding business activities. Such efforts would be linked with their efficiency in asset utilization to generate revenues, indicating lower capital intensity. Additionally, firms should continually innovate their traditional working process and apply advanced technology to adapt to rapid environmental changes and compete with their rivals. Last but not least, it is suggested that SMEs' managers update their knowledge and follow professional training or education programs to have more qualifications and readiness to learn technology, initiate innovation and develop new products/services for better firms' competitiveness.

## References

- Aghion, P., Howitt, P.: A model of growth through creative destruction. *Econometrica* **60**, 323–351 (1992)
- Agnew, J., Balduzzi, P., Sunden, A.: Portfolio choice and trading in a large 401 (k) plan. *Am. Econ. Rev.* 193–215 (2003)
- Andrew, W.D., Damitio, J.W., Schmidgall, R.S.: *Financial Management for the Hospitality Industry*. Pearson Prentice Hall, Upper Saddle River (2007)
- Baldwin, S.A., Fellingham, G.W.: Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychol. Methods* **18**(2), 151 (2013)
- Barber, B.M., Odean, T.: Boys will be boys: gender, over confidence, and common stock investment. *Q. J. Econ.* 261–292 (2001)
- Børing, P.: Effect of firms' age on their use of highly skilled workers. *Labour* **34**(2), 137–153 (2020)
- Brooks, S., Gelman, A.: General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**(4), 434–455 (1998)
- Chen, H., Volpe, R.P.: Gender differences in personal financial literacy among college students. *Financ. Serv. Rev.* **11**(3), 289–307 (2002)
- Crosno, R., Gneezy, U.: Gender differences in preferences. *J. Econ. Lit.* 448–474 (2009)
- Eckel, C.C., Grossman, P. J.: Differences in the economic decisions of men and women: experimental evidence. In: *Handbook of Experimental Economics Results*, vol. 1, pp. 509–519 (2008)

- Epstein, S.: Cognitive-experiential self-theory of personality. In: Handbook of Psychology (2003) Finance Management: Capital Intensity Ratio – Meaning, Formula, Importance, and More. Finance Management (2021). <https://efinancemanagement.com/financial-analysis/capital-intensity-ratio>
- Fonseca, R., Mullen, K.J., Zamorro, G., Zissimopoulos, J.: What explains the gender gap in financial literacy? The role of household decision making. *J. Consum. Aff.* **46**(1), 90–106 (2012)
- Frankenfield, J.: What is Capital Intensive? Investopedia (2020). <https://www.investopedia.com/terms/c/capitalintensive.asp>
- Gelman, A., Rubin, D.B.: Inference from Iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992)
- Grossman, G., Helpman, E.: Innovation and Growth in the Global Economy. MIT Press, Cambridge (1991)
- Hira, T.K., Mugenda, O.: Gender differences in financial perceptions, behaviors and satisfaction. *J. Financ. Plan.-Denver-* **13**(2), 86–93 (2000)
- Hurdle, G.: Leverage risk market structure, and profitability. *Rev. Econ. Stat.* **56**, 478–485 (1974)
- Judzik, D., Sala, H.: The determinants of capital intensity in Japan and the US. *J. Jpn. Int. Econ.* **35**, 78–98 (2015)
- Lee, S.: Effects of capital intensity on firm performance: the US restaurant industry. *J. Hosp. Financ. Manag.* **18**(1), 1–13 (2010)
- Lubatkin, M., Chatterjee, S.: Extending modern portfolio theory into the domain of corporate diversification: does it apply? *Acad. Manag. J.* **37**(1), 109–136 (1994)
- Lusardi, A., Mitchell, O.S.: Planning and Financial Literacy: How Do Women Fare? (No. w13750). National Bureau of Economic Research (2008)
- Lusardi, A., Mitchell, O.S.: Financial literacy around the world: an overview. *J. Pension Econ. Finance* **10**(04), 497–508 (2011)
- Muzakki, M.R., Darsono, D.: The influence of corporate social responsibility and capital intensity on tax avoidance. *Diponegoro J. Account.* **4**(3), 1–8 (2015)
- Ni, B.: Productivity, Capital Intensity and ISO14001 Adoption-Theory and Evidence from Vietnam. Graduate School of Economics and Osaka School of International Public Policy (OSIPP). Osaka University Discussion Papers in Economics and Business, vol. 15, pp. 1–24 (2015)
- Penny, W.D., Mattout, J., Trujillo-Barreto, N.: Bayesian model selection and averaging. In: Statistical Parametric Mapping, pp. 454–467 (2007). <https://doi.org/10.1016/b978-012372560-8/50035-8>
- Queiró, F.: The effect of manager education on firm growth. *QJ Econ.* **118**(4), 1169–1208 (2016)
- Romer, P.M.: Endogenous technological change. *J. Polit. Econ.* **98**(5), 1–102 (1990)
- Shojaie, T., Tehranchian, A.M.: New empirical evidence on the determinants of capital intensity: an adaptive comparison of Iran and China. *Int. J. Econ. Financ. Issues* **8**(2), 94 (2018)
- Sladek, R.M., Bond, M.J., Phillips, P.A.: Age and gender differences in preferences for rational and experiential thinking. *Personal. Individ. Differ.* **49**(8), 907–911 (2010)
- Söderbom, M., Teal, F.: Firm Size and Human Capital as Determinants of Productivity and Earnings (2001)
- Urrahmah, S., Mukti, A.H.: The effect of liquidity, capital intensity, and inventory intensity on tax avoidance. *Target* **2017**(2018), 2019 (2016)
- Dao, L.V.P.: Analyzing the differences in the impact of FDI and exports on labor productivity of enterprises: the case of Vietnam. In: Nguyen, A.T., Hens, L. (eds.) Global Changes and Sustainable Development in Asian Emerging Market Economies Vol. 1, pp. 211–224. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-81435-9\\_16](https://doi.org/10.1007/978-3-030-81435-9_16)
- Webster, R.L., Ellis, T.S.: Men's and women's self-confidence in performing financial analysis. *Psychol. Rep.* **79**(3f), 1251–1254 (1996)
- Zissimopoulos, J.M., Karney, B., Rauer, A.: Marital Histories and Economic Well-Being. Michigan Retirement Research Center Research 180 (2008)



# Labor Productivity: Does Export Matter for Vietnamese Small and Medium Enterprises?

Dang Nhan Truong Thanh, Ha Van Dung, and Nguyen Van Tung<sup>(✉)</sup>

Banking University, Hochiminh City, Vietnam  
{nhandtt, dunghv, tungnv}@buh.edu.vn

**Abstract.** Several micro-data studies have found one-directional effect in the relationship between productivity and exports. Nevertheless, there is little evidence of the significant impacts of exports on productivity. Particularly, there is a literature gap regarding assessing whether engaging in exporting indeed influences productivity at the firm level in developing countries such as Vietnam. With an emphasis on Vietnam, this study aims to respond to the demand for further empirical evidence on the influence of exports on labor productivity. The research's objective is to analyze the influence of exporting on labor productivity by examining Small and Medium Enterprises (SMEs) across different industries in the country's context. According to the empirical findings using Bayesian statistics, export-oriented SMEs in Vietnam saw greater labor productivity. With managerial implications, firms need to have continual innovations in terms of technology, process, and products. They also need to keep learning from foreign markets in terms of buyer-seller relationships, quality management, foreign suppliers, international supply chain, foreign trade, technological innovation, and regulatory issues.

**Keywords:** labor productivity · export · SMEs · Vietnam · Bayesian estimation

**JEL Classification:** B26 · G21 · C11

## 1 Introduction

There has been ongoing controversy regarding whether policies of export promotion or import substitution offer the best economic results for a country (Yasar and Nelson, 2003). This debate resulted in some new evidence proving that there is a correlation between exporting and economic performance (Havrylyshyn, 1990; Feenstra, 1995; Edwards, 1998), and encouraged some developing countries to switch from import substitution to outward- and market-oriented policies in the 1970s and 1980s (Yasar and Nelson, 2003).

Even though the role of exports in stimulating growth in general, and productivity in particular, has been investigated empirically using aggregate data for countries and industries for a long time, only recently have there been more studies at the firm level looking at the extent and causes of productivity differentials between exporters and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 274–285, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_19](https://doi.org/10.1007/978-3-031-35763-3_19)

their counterparts which only provide products within the domestic market (Schwarzer, 2017).

In practice, business productivity levels undeniably depend on business decisions and can change over time. Similarly, entry into and exit from exports is a recurring feature of individual firms. In recent years, there has been growing interest in better understanding the causal direction of the strong correlation between productivity and exporting (Schwarzer, 2017).

Several micro-data studies have found a directional correlation between productivity and exports, indicating that productivity increases exports. Nevertheless, there is little evidence of the significant impacts of exports on productivity. The results from previous studies on the causal link between exports and productivity at the plant level suggest a need for further investigation across different industries and within different contexts (Yasar, 2013).

Following economic collapse in the mid-1980s, Vietnam initiated its Doi Moi 'renovation' reform process in 1986. A highly coordinated set of public investments, targeted policies, and institutional initiatives was applied with the aim of facilitating trade and promoting exports, much like the export efforts in other East Asian countries in the 1970s. WTO membership followed later in 2007.

The aggregate growth rates have been strikingly high in Vietnam since 1986 when the country embarked on respectively a standard package of stabilization and structural adjustment (Arndt et al., 2000) and Doi Moi (Arndt et al., 2012). This resulted in a completely changed landscape for industrial development, increased trading opportunities and a drastic fall in poverty headcount rates (Newman et al., 2017). Structural transformation remains sluggish, and the enterprise sector continues to struggle to survive with the elusive goal of breaking into export markets (Jones and Tarp, 2013; Newman et al., 2017).

There is a need to carefully consider whether engaging in exporting influences productivity at the firm level in developing countries such as Vietnam (Newman et al., 2017). In the comprehensive review of the literature, Syverson (2011) cautioned that despite the widely identified strong correlation between the average productivity level of an industry's plants and that industry's trade exposure, there seems to be less evidence of large productivity influences on the domestic plants when they begin exporting. Many studies have found that this correlation largely reflects selection rather than a causal impact of exporting on productivity (Newman et al., 2017). Based on these discussions, this research aims to respond to the need for more empirical evidence about the impact of exporting on labor productivity with a focus on Vietnam. The research's objective is to analyze the influence of exporting on labor productivity by examining Small and Medium Enterprises (SMEs) across different industries in the country's context. Based on these results, several managerial implications would be provided to support firms in executing more effective exporting activities and enhance better labor productivity.

On the theoretical aspect, this research can contribute to the literature development about the topic on enterprises' labor productivity and exporting, focusing on SMEs in a developing country such as Vietnam. This empirical study can be viewed as a response to the call raised by previous researchers for more studies on the causal relationship between exporting and labor productivity in Vietnam (Jones and Tarp, 2013; Newman et al., 2017)

as well as a considerable reference for academics who have interest in labor productivity in emerging countries. Practically, this study can help to provide managerial implications for the purpose of enabling firms' labor productivity and eventually contributing to economic growth.

## 2 Literature Review

There have been various definitions of labor productivity among different countries and industries. In every field or industry sector, modifications, specifications or levels of details focused on particular needs to come up with new measures of labor productivity can be employed (Bureš and Stropková, 2014). Mohanty (1992) provided 12 distinct definitions of productivity classifying definitions as macro-level and micro-level. Bernolak (1997) and Hannula (2002) discussed the use and applicability of various methods at length.

Enshassi et al. (2007) defined *productivity* as the ratio of outputs to inputs. Durdyev et al. (2012), meanwhile, conceptualized productivity as the quantity of work produced per man-hour,

equipment-hour, or crew-hour worked. According to Nasirzadeh and Nojedehei (2013), *labor productivity* is defined as the ratio between completed work and expended work hours to execute the project. Goel and Agraewal (2017), synthesized the main criteria used in defining productivity, including: "effective utilization of resources", "well defined and clear objectives" and "resources contributing to the achievement of desired objectives".

### Theoretical Background Related to the Correlation Between Exporting and Labor Productivity

Models advocated by Krugman (1979) and Jovanovic and Lach (1991) demonstrated that exporting can enhance labor productivity by means of the following: (1) exporting firms learn about and adopt the best international production, distribution and management methods, (2) exporting firms receive feedbacks from international customers, suppliers and competitors that can help to improve their product standards as well as benefit from other knowledge spillovers (Yasar, 2013).

In the work-related theory of international trade, developed by Melitz (2003), firms are endowed with differential productivity gains that are predetermined and invariant over time. Only firms that achieve productivity gains above the export threshold are likely to enter foreign markets. In fact, this prediction has been broadly supported by empirical research and corporate performance. Exports are, on average, more productive than purely domestic operating counterparts (Bernard et al., 2012).

Two hypotheses can be applied as theoretical background which help to explain the mechanism underlying the "black box" of higher observed productivity in exporting firms:

*Self-selection and Learning-By-Exporting (LBE).*

The *self-selection* hypothesis implies that firms becoming exporters are simply more productive to start with. This hypothesis suggests that firms with higher productivity

“self-select” into exporting, as their productivity edge allows them to pay off the higher costs of serving foreign markets. There is a broad consensus in the empirical literature reviewed in Wagner (2007), Greenaway and Kneller (2007a), and Bernard et al. (2012). These researchers also highlighted substantial differences in firm-level productivity between domestically operating firms and future exporters prior to their entry into exporting.

The *Learning-By-Exporting (LBE)* hypothesis states that firms increase their productivity as a result of exporting. The early construction of this hypothesis can be traced back to endogenous growth models, such as Grossman and Helpman (1993), who discussed technology diffusion through participation in international markets which enhanced internal productivity. Demand-side-driven exploitation of economies of scale was emphasized as an important productivity-causing factor in traditional export-led growth hypotheses (Kaldor, 1970). LBE hypotheses highlight a variety of mechanisms such as learning from foreign markets in terms of buyer-seller relationships, and increased competition with foreign suppliers, or adapting and improving product quality to suit foreign preferences, which lead to the improvement in labor productivity (Schwarzer, 2017).

### **Previous Studies Related to the Correlation Between Exporting and Labor Productivity**

Several empirical studies generally remain doubtful about the precise mechanisms underlying LBE, but generally considered that productivity effects of firms' international activities are, by definition, inherent in entering a foreign market, entailing activities and knowledge that non-exporters do not acquire. The evidence for this effect has been provided by previous studies such as Hosono et al. (2015), Fernandes and Isgut (2015), Manjon et al. (2013), Lileeva and Trefler (2010), Bigsten and Gebreeyesus (2009), De Loecker (2007), Van Biesebroeck (2005) and Girma et al. (2004).

The study of Ahn (2005) explored a plausible channel through which exporting could have created both a substantial and a persistent contribution to productivity growth in a developing economy and export-oriented economic growth in Korea. The aim of this paper was to review recent empirical findings related to the connection between exporting and productivity to offer some new evidence from Korean microdata. Plant-level data for Korean manufacturing demonstrated that more export-intensive industries tend to have a higher productivity level. Additionally, a substantial part of the variance in plant level productivity is caused by the variance in industry-level export intensity.

Verhoogen (2008) concluded that the decision to export was related to product quality upgrading for the objective of serving consumer preferences in the foreign market, leading to higher productivity.



Aw et al. (2011) proposed a model of endogenous R&D decisions jointly with exporting, which can explain the post-entry productivity growth of exporters. Later, Albornoz et al. (2012) developed a model which assumes uncertainty about firms' general ability to earn profits abroad, which might be solved only through trial-and-error experience in foreign markets.

Bustos (2011) built a model of technology adoption jointly with entry to exporting and found empirical evidence for trade liberalization, which induced innovation in both new and existing exporters from Argentina. Importantly, this mechanism seems to be generalizable in advanced economies: Lileeva and Trefler (2010) found evidence for similar predictions for Canadian exporters following tariff reductions in the US.

Mayer et al. (2014) discussed that multi-product firms point to adjustments in product mixes due to increased competition in export markets, which encourages their focus on their core competencies and adjusts their product offer accordingly, resulting in firm-level productivity gains.

Based on the application of a very comprehensive dataset on Danish firms in services and manufacturing, Malchow-Møller et al. (2015) separated services and goods traders and investigated the respective links with long-term (2002 - 2008) productivity growth. Their findings suggested that organizations that started exporting goods in this period experienced higher average productivity growth than firms that had never exported in this period.

All these papers underlined a specific aspect that may help to reason the observed correlation between exporting and productivity. However, the theoretical basis for motivating either effect is often controversial, as the models of dissimilar firms are generally static (Schwarzer, 2017). According to Tybout (2003), the recognition of the immediate link between productivity at the time of entry into exporting is usually problematic, as the econometrician normally does not have all the necessary information, especially around the time of entry into exporting. In other words, the decision to enter exporting may be more relevant than the actual entry into exporting.

### **Research Hypotheses**

Based on the models of Krugman (1979) and Jovanovic and Lach (1991), the theoretical background of Self-Selection and Learning-By-Exporting (LBE) and previous studies such as Hosono et al. (2015), Malchow-Møller et al. (2015), Mayer et al. (2014), Fernandes and Isgut (2015), Manjon et al. (2013), Aw et al. (2011), Lileeva and Trefler (2010), Bigsten and Gebreeyesus (2009), De Loecker (2007), Van Biesebroeck (2005) and Girma et al. (2004), this study aimed to investigate the effects of some factors including exporting, firm size, firm age and managers' gender on productivity.

## **3 Methodology**

The paper uses a dataset of more than 2,500 small and medium enterprises in Vietnam from 2007 to 2015 compiled by the Central Institute for Economic Management (CIEM), the Institute of Labor Science and Social Affairs (ILSSA), the United Nations University's Institute for World Development Economics (UNU-WIDER) and the Faculty of Economics (DOE) of the University of Copenhagen jointly investigated the research.



Based on the stated hypotheses, the article introduces the following variables into the model:  $\ln LP_{it} = \beta_1 + \beta_2 \ln TL_{it} + \beta_3 \ln equity_{it} + \beta_4 ex_{it} + \beta_5 y\_est_{it} + \beta_6 gender_{it} + \beta_7 z1_i + \beta_8 z2_i + \beta_9 z3_i + \varepsilon_i$

**Where:**

**Dependent Variable**

LP is the labor productivity is the rate of total revenue over total employees.

**Explanatory Variable**

Ex is a binary variable, representing the export factor. If firms export,  $ex = 1$  and otherwise 0.

**Control Variables**

TL is total full-time employees, measured in natural logarithm.

Equity is total equity at the end of the fiscal year, measured in natural logarithms.

y\_est is number of years of firm establishment, which is used for firm age.

gender is the manager gender, equals 1 if firm manager is male, otherwise 0.

z1, z2, and z3 are dummy variables for manager educational level. z1 equals 1 if the manager graduated from university, and otherwise z1 is 0. z2 equals 1 if the manager graduated from vocational training, and otherwise 0. z3 receives the value of 1 if the manager got training without degree, and otherwise 0.

The Bayesian estimation method is used to estimate the model parameters because of its advantages. Bayesian inference is based on a single probabilistic rule – the Bayesian rule, which is applied to all parametric models (Berger, 1993), thus making the Bayesian approach universal and facilitating significant advantages for the application and interpretation of the results. Bayesian combination of a priori information and data, gives more reliable estimates than other methods in the case of small samples, lacking data (Baldwin & Fellingham, 2013).

## 4 The Results

In order to select the appropriate a priori information for large sample size, the article will analyze the sensitivity through five simulations of normally distributed a priori information as follows (Table 1):

**Table 1.** Likelihood model

	lnLP $N(\mu, \delta)$
Prior distributions:	
Simulation 1	$\alpha_i \sim N(0, 1)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 2	$\alpha_i \sim N(0, 10)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha_i \sim N(0, 100)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01) S$
Simulation 4	$\alpha_i \sim N(0, 1000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha_i \sim N(0, 10000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
$i = 1, 2, 3, 4, 5$	

To select the most suitable simulation, Bayesian coefficients and Bayesian estimation models will be performed based on Log BF, Log (ML) and DIC criteria. In which, the selected criteria will be the maximum mean for Log BF, Log (ML), P(M/y), and the smallest for DIC (Penny et al., 2007).

**Table 2.** Bayesian factor test and model test

	Chan	lnLP			
		Avg DIC	Avg log (ML)	Avg log (BF)	P(M/y)
Simulation 1	3	2.91e+04	-1.46e+03		0.0000
Simulation 2	3	2.91e+04	-1.46e+03	<b>38.4527</b>	<b>0.9957</b>
Simulation 3	3	2.91e+04	-1.46e+03	32.9993	0.0043
Simulation 4	3	2.91e+04	-1.46e+03	22.9178	0.0000
Simulation 3	3	2.91e+04	-1.46e+03	12.5683	0.0000

Source: Author’s calculation

Based on the results of Table 2, Model 2 is selected.

After the model is selected, the MCMC convergence needs to be considered. According to Gelman and Rubin (1992), Brooks and Gelman (1998), diagnostic Rc values greater than 1.2 for any model parameter are considered non-convergent. In practice,  $Rc < 1.1$  is often used to declare convergence. Therefore, the table above has a Max Gelman-Rubin Rc value  $< 1.1$  indicating that the MCMC convergence is acceptable for Bayesian analysis (Table 3).

**Table 3.** Bayesian simulation results

lnLP						
	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
lnTL	-.2429999	.0115823	.000067	-.2430225	-.2656411	-.2202605
Inequity	.2938167	.0069817	.00004	.2938724	.2801075	.3074261
ex	.2327238	.0419716	.000243	.2327143	.1504462	.3149633
y_est	-.0110262	.0008968	5.2e-06	-.011034	-.0127789	-.0092773
gender	-.0623961	.0191846	.000111	-.062394	-.1003088	-.024774
z1	.3453539	.0312218	.000181	.3451359	.2842252	.4060998
z2	.2807738	.027176	.000157	.2807692	.2277706	.3342599
z3	.0809437	.0260722	.000151	.0810188	.030007	.1322228
_cons	10.38535	.0462014	.000267	10.38569	10.29455	10.4755
var	.8989753	.0123631	.000072	.8989147	.8748719	.9234982

Number of obs = 10,654

Avg acceptance rate = 1

Avg efficiency: min = .98

Max Gelman-Rubin Rc = 1

Source: Author's calculation

To visually check the MCMC convergence of lnLP, the author used histograms to consider the degree of autocorrelation, normal distribution, and stability. The resulting histograms show low autocorrelation, while the trace plots show a good association. Normal distributions can be plotted from density histograms and frequency distribution histograms. Therefore, it can be concluded that MCMC is convergent.

The author also calculated to evaluate the sensitivity. The results of the certainty test show that the posterior estimates are not significantly different in terms of the posterior mean. MCSE and confidence intervals when the normal base values for all parameters are adjusted from -0.5 to 0.5 with a 0.1 interval. Therefore, it can be said that the Bayesian inference results are reasonable, and the estimated model is stable.

The Bayesian results show some similarities findings of this research to previous ones. Firstly, the findings strongly support the Self-Selection and Learning-By-Exporting theory as well as earlier studies (Wagner, 2007; Greenaway and Kneller, 2007a; Bernard et al., 2012; Mayer et al., 2014; Hosono et al., 2015; Fernandes and Isgut, 2015). The finding indicate that exporting activity leads to higher firms' productivity. It seems that to export products to foreign markets, firms have to innovate technologies for their higher competition. In the case of developing economies like Vietnam, the most important markets for exported products are developed economies, which demand high product quality (Ahn, 2005). As a result, demand for high productivity rises. When a firm wants to export its products, it has to prepare for competition by capital and technology. The competition may come from host firms or other exporters. Capital preparation may help

firms to cover different fees such as new market analysis, advertisement, promotion, etc. One of the strategies to get into the export market is low prices. This price level will exist for a period, which leads to pressure on firms to raise their productivity as the spillovers of Learning-By-Exporting in Vietnamese SMEs. The same impacts can be found in technology. High demand for product quality requires a technology/capital-intensive production process and Self-Selection is proven.

Secondly, bigger firms are assumed to have a larger ability to absorb capital and innovative technology, which in turn increases productivity (Malchow-Møller et al., 2015; Mayer et al., 2014; Fernandes and Isgut, 2015). The results of this paper partly confirm that bigger firms will have higher labor productivity. In the same direction as most previous studies, bigger firms measured by firm equity have higher productivity. Firm equity is part of total assets formed by firm physical capital. For SMEs in developing economies, the investment in physical capital is quite limited; thus, the technological advance in capital-intensive firms is not far ahead of others. However, the investment in physical capital is found to have positive impacts on labor productivity in this research. On the other hand, firm size measured by total employment indicates a negative effect on productivity. Most domestic firms in developing economies are labor-intensive ones. Developing from the low technology economy and skipping capitalism, the labor force in Vietnam cannot immediately fulfill the skill requirements of a developed economy. The heavy industry-based economy was already a failure in the past in Vietnam and now the processing industry and light industry are the base for Vietnam's economic development. Bigger firms absorbing larger labor with low-skill, low-disciplined workers, are more difficult to increase their productivity than small and medium firms. The advantages of SMEs are that they can upgrade their technology for a relatively small system which costs less time and capital. SMEs can also more easily replace a part of their production process without affecting other production procedures. The smaller are the firms, the more flexible they are.

Thirdly, younger firms are empirically found to have higher labor productivity. This means labor productivity will be lower in older firms. The results are similar to those of the previous (Bigsten and Gebreeyesus, 2009; De Loecker, 2007). To compete internationally, firms must focus on their self-selection technology. One of the advantages of younger firms is that these firms have more chances to choose up-to-date and suitable technology, which is relatively high capital intensive. Modern technology will help younger firms raise their labor productivity.

Last but not least, gender of managers has positive impacts on labor productivity. This finding is consistent with previous studies that labor productivity is higher in firms with male managers (Islam et al., 2020). Both constraint and preference perspectives could be held in developing economies. The constraint perspective implies that gender differences would discourage women in management. Due to broader social relationships, men may have more opportunities to access finance and procedure. The preference perspective indicates that gender-specifics give different business approaches. Female managers often consider the low productivity sector in order to balance time to work and time for family. A risk-averse manner could be another explanation for the low labor productivity in female-managed firms.

## 5 Conclusion

More than 2,500 small and medium enterprises in Vietnam from 2007 to 2015 are used for the analysis of the role of the export factor on labor productivity. The empirical results with Bayesian statistics show that export-based SMEs in Vietnam experienced higher labor productivity. These results are consistent with what has been mentioned in literature as well as in previous studies. The export-oriented policy is supported in this case at the micro level of SMEs. Additionally, bigger firms measured by firm equity have higher productivity, which is similar to most previous studies. Besides, younger firms are empirically found to have higher labor productivity. Last but not least, the gender of managers has positive impacts on labor productivity.

Based on the application of the self-selection and Learning-By-Exporting hypothesis as well as research findings, it is highly recommended that SMEs in Vietnam should invest more in exporting activities which would lead to higher labor productivity. In order to be more ready for exporting practices and taking more advantages of such practices, firms need to have continual innovations in terms of technology, process, and products. They also need to keep learning from foreign markets in terms of buyer-seller relationships, quality management, foreign suppliers, international supply chain, foreign trade, technological innovation, and regulatory issues. From the government side, there should be clearer instructions and policies regarding the allowance of exporting.

Theoretically, this research can contribute to the literature development in the field of firms' labor productivity and exporting, focusing on SMEs in a developing country such as Vietnam. This empirical study can be considered as a response to the call raised by previous researchers for more studies on the causal relationship between exporting and labor productivity in Vietnam (Jones and Tarp, 2013; Newman et al., 2017) as well as a reference for future related studies. Practically, this study provides managerial implications for the improvement of firms' labor productivity and eventually fostering economic growth.

## References

- Ahn, S.: Does Exporting Raise Productivity? Evidence from Korean Microdata. ADB Institute Research Paper Series no. 67 (2005)
- Arndt, C., Garcia, A.F., Tarp, F., Thurlow, J.: Poverty reduction and economic structure: comparative path analysis for Mozambique and Vietnam. *Rev. Income Wealth* **58**(4), 742–763 (2012)
- Arndt, C., Jensen, H.T., Tarp, F.: Stabilization and structural adjustment in Mozambique: an appraisal. *J. Int. Dev.* **12**(3), 299–323 (2000)
- Aw, B.Y., Roberts, M.J., Xu, D.Y.: R&D Investment, exporting, and productivity dynamics. *Am. Econ. Rev.* **101**(4), 1312–1344 (2011)
- Baldwin, A., Fellingham, W.: Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychol. Methods* **18**(2), 151–164 (2013)
- Berger, J.O.: *Statistical Decision Theory and Bayesian Analysis*. Springer, New York (1993)
- Bernard, A.B., Jensen, J.B., Redding, S.J., Schott, P.K.: The empirics of Firm Heterogeneity and International Trade. *Ann. Rev. Econ.* **4**, 283–313 (2012)
- Bernolak, I.: Effective measurement and successful elements of company productivity: the basis of competitiveness and world prosperity. *Int. J. Prod. Econ.* **52**(1), 203–213 (1997)

- Bigsten, A., Gebreeyesus, M.: Firm productivity and exports: evidence from Ethiopian manufacturing. *J. Dev. Stud.* **45**(10), 1594–1614 (2009)
- Brook, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**(4), 434–455 (1998)
- Bureš, V., Stropková, A.: Labour productivity and possibilities of its extension by knowledge management aspects. *Procedia Soc. Behav. Sci.* **109**, 1088–1093 (2014)
- Bustos, P.: Trade liberalization, exports, and technology upgrading: evidence on the impact of Mercosur on Argentinian firms. *Am. Econ. Rev.* **101**(1), 304–340 (2011)
- De Loecker, J.: Do exports generate higher productivity? evidence from Slovenia. *J. Int. Econ.* **73**(1), 69–98 (2007)
- Edwards, S.: Openness, productivity and growth: what do we really know? *Econ. J.* **108**, 383–398 (1998)
- Enshassi, A., Mohamed, S., Mustafa, Z.A., Mayer, P.E.: Factors affecting labour productivity in building projects in the gaza strip. *J. Civ. Eng. Manag.* **13**(4), 245–254 (2007)
- Feenstra, R.: Estimating the effects of trade policy. In: Grossman, G., Rogoff, K (eds.) *Handbook of International Economics*, vol. 3, pp. 1553–1595 (1995)
- Fernandes, A.M., Isgut, A.E.: Learning-by-exporting effects: are they for real? *Emerg. Mark. Financ. Trade* **51**(1), 65–89 (2015)
- Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**(4), 457–511 (1992)
- Girma, S., Greenaway, D., Kneller, R.: Does exporting increase productivity? a microeconomic analysis of matched firms. *Rev. Int. Econ.* **12**(5), 855–866 (2004)
- Goel, V., Agrawal, R.: Factor affecting labour productivity: an integrative synthesis and productivity modelling. *Global Bus. Econ. Rev.* **19**(3), 299–322 (2017)
- Grossman, G.M., Helpman, E.: *Innovation and Growth in the Global Economy*. MIT press (1993)
- Hannula, M.: Total productivity measurement based on partial productivity ratios. *Int. J. Prod. Econ.* **78**(1), 57–67 (2002)
- Havrylyshyn, O.: Trade policy and productivity gains in developing countries: a survey of the literature. *World Bank Res. Obs.* **5**, 1–24 (1990)
- Hosono, K., Miyakawa, D., Takizawa, M.: Learning by Export: Does the Presence of Foreign Affiliate Companies Matter? Technical Report, RIETI Discussion Paper, 15-E-053 (2015)
- Islam, A., Gaddis, I., Lopez, A.P., Amin, M.: The labor productivity gap between formal businesses run by women and men. *Fem. Econ.* **26**(4), 228–258 (2022). <https://doi.org/10.1080/13545701.2020.1797139>
- Jones, S., Tarp, F.: Jobs and welfare in Mozambique. In: WIDER Working Paper 2013/045, UNU-WIDER background paper to the 2012 World Development Report (2013)
- Jovanovic, B., Lach, S.: The diffusion of technology and inequality among nations. NBER Working Paper (1991)
- Kaldor, N.: The case for regional policies. *Scottish J. Polit. Econ.* **17**(3), 337–348 (1970)
- Krugman, P.: A model of innovation, technology transfer, and the world distribution of income. *J. Polit. Econ.* **87**, 253–266 (1979)
- Lileeva, A., Trefler, D.: Improved access to foreign markets raises plant-level productivity for some plants. *Q. J. Econ.* **125**(3), 1051–1099 (2010)
- Malchow-Møller, N., Munch, J.R., Skaksen, J.R.: Servicetrade, goods trade and productivity growth: evidence from a population of private sector firms. *Rev. World Econ.* **151**(2), 197–229 (2015)
- Manjon, M., Manez, J.A., Rochina-Barrachina, M.E., Sanchis-Llopis, J.A.: Reconsidering learning by exporting. *Rev. World Econ.* **149**(1), 5–22 (2013)
- Mayer, T., Melitz, M.J., Ottaviano, G.I.P.: Market size, competition, and the product mix of exporters. *Am. Econ. Rev.* **104**(2), 495–536 (2014)

- Melitz, M.J.: The impact of trade on intra-industry real locations and aggregate industry productivity. *Econometrica* **71**, 1695–1725 (2003)
- Mohanty, R.P.: Consensus and conflicts in understanding productivity. *Int. J. Prod. Econ.* **28**(1), 95–106 (1992)
- Nasirzadeh, F., Nojehdehi, P.: Dynamic modeling of labor productivity in construction projects. *Int. J. Project Manage.* **31**(6), 903–911 (2013)
- Newman, C., Rand, J., Tarp, F., Anh, N.T.T.: Exporting and productivity: learning from Vietnam. *J. Afr. Econ.* **26**(1), 67–92 (2017)
- Penny, W. D., Mattout, J., Trujillo-Barreto, N.: Bayesian model selection and averaging. In: *Statistical Parametric Mapping*, pp. 454–467 (2007). <https://doi.org/10.1016/b978-012372560-8/50035-8>
- Schwarzer, J.: The effects of exporting on labor productivity: evidence from german firms. CEP Working Paper, 2 (2017)
- Syversen, C.: What determines productivity? *J. Econ. Lit.* **49**(2), 326–365 (2011)
- Tybout, J.R.: Plant-and firm-level evidence on “new” trade theories. In: *Handbook of International Trade*, vol. 1, pp. 388–415 (2003)
- Van Biesebroeck, J.: Exporting raises productivity in sub-saharan African manufacturing firms. *J. Int. Econ.* **67**(2), 373–391 (2005). <https://doi.org/10.1016/j.jinteco.2004.12.002>
- Verhoogen, E.A.: Trade, quality upgrading, and wage inequality in the mexican manufacturing sector. *Q. J. Econ.* **123**(2), 489–530 (2008)
- Wagner, J.: Exports and productivity: a survey of the evidence from firm-level data. *The World Economy* **30**(1), 60–82 (2007)
- Yasar, M.: Direct and indirect exporting and productivity: evidence from firm-level data. *Manag. Decis. Econ.* **36**, 109–120 (2013)
- Yasar, M., Nelson, C.H.: The Relationship between Exports and Productivity at the Plant-level in the Turkish Apparel and Motor Vehicle Parts Industries. Unpublished Manuscript Emory University. Home Foreign (2003)



# Income and Consumption Patterns of Sri Lankan Senior Citizens and Subsequent Impact on Policies and Transportation

Shanika Madushani Jayathunga<sup>1,2</sup> and Gnanadarsha Sanjaya Dissanayake<sup>2,3</sup> (✉)

<sup>1</sup> University of Colombo, Colombo, Sri Lanka

<sup>2</sup> School of Mathematics and Statistics, University of Sydney, Camperdown, Sri Lanka  
gnanadarshad@gmail.com

<sup>3</sup> NSW Ministry of Health, St. Leonards, Australia

**Abstract.** Sri Lanka's population is continuously growing older. Some studies show the elderly population or senior citizens of Sri Lanka will rapidly grow to 30% with an increase in the dependency ratio by 2060.

However, there are a considerable number of elders, who are still working and contributes to the country's economy. The elders have different spending patterns depending on their level of income. Identifying the patterns of elderly income and expenditure are important in arriving at government policy decisions. Embedded within such initiatives will be policies targeting to address the continuance of social welfare catering especially to the needs of the elderly population.

This paper focusses on identifying sources of elderly income and expenditure patterns in Sri Lanka and the resulting impact contribution towards the labour force.

**Keywords:** Elders · Income · Expenditure · Labour · Policy · Economy · Accessibility

## 1 Introduction and Literature Review

Income distribution or the distribution of income is a statistical measure of how many people earn or receive various amounts of income. In terms of economics an income distribution is defined as how a nation's total Gross Domestic Products (GDP) is distributed amongst its population. *Income distribution patterns or income patterns* on the other hand are diverse across countries and unrelated in any obvious way to per capita GDP or the rate of GDP growth.

*Consumption pattern* describes how consumers act, how they allocate income among various alternatives, how loyal they are to various brands and how they react to new products and services. In economics it has been mostly used to characterize a household's allocation of expenditures across different consumption categories, such as food, housing, clothing, and transportation.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSSC 483, pp. 286–298, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_20](https://doi.org/10.1007/978-3-031-35763-3_20)



In the current context, elderly citizens are expected to remain independent and active and use transport systems frequently until they reach their eighties in terms of age (Siren and Haustein, 2013). Developing countries are moving towards legislation that requires transport services to be made accessible, although the translation of accessibility policies into the provision of inclusive transport remains a major obstacle for many reasons, e.g., lack of monitoring and enforcement of compliance with existing accessibility legislation and inadequate resources for implementation (Roberts and Babinard, 2004).

The word Elderly differ from country to country. In Western nations such as USA, Australia etc., the elderly population cohort is referred to as “Senior Citizens”. In such nation’s individuals above the age of 60 fall into the elderly segment, when considered for retirement benefits and other social welfare schemes. On the contrary in a generic sense, European countries consider age of 65 as the demarcation cut-off measure as opposite to the developing world conventional age of 60 and above considered as the elderly segment (Menike, 2014). In Sri Lanka, age 55 and above are considered as elderly since, it becomes the retirement age of most private and public sector organizations. Sri Lanka has entered the third stage of demographic transition (Attanayake, 1984) and due to that Sri Lanka has a very high elderly population among south Asian countries forming 13.2% of the total population in 2001. It is predicted to reach 20% by 2020 according to the Department of Census and Statistics in Sri Lanka.

However, Sri Lanka has a culture where the younger people look after their elderly relatives, since most of the senior citizen population are reluctant to work beyond retirement. But senior citizens in developed countries work beyond the Sri Lankan retirement age threshold due to their own disposable income savings, lifetime investments and retirement pension plans. In such a context, Australia can be presented as a classic example. It is a country that has a retirement scheme based on the date of birth of each individual worker pertaining to a minimum cut-off age of retirement and the ability that an able-bodied Australian could work until they feel they are capable. But in Sri Lanka with a rising proportion of the elderly segment, the dependency burden has been rising among the working age population (Mendis, 2007). Dependency ratio is defined as the number of dependents per 100 people in the working age group (age 15–59 referred to as prime-age adults) (De Silva, 2012). For obvious reasons children (individuals under age 15) and older persons (those aged 60 and above) are considered to be dependent. At present, there are 40 children and 20 senior citizens per every 100 prime-age adults in the country. By 2060, the old age dependency ratio is estimated to increase close to 50 and the child dependency ratio is estimated to decline to 30 per every 100 prime-age adults (United Nation Economic and Social Commission for Asia and the Pacific-2015). According to De Silva (2008) the aging in Sri Lanka seems to occur at lower levels of economic development compared with those of western countries.

The expenditure pattern varies among working and non-working aging populations due to home-made meals, health insurance, social security and retirement contributions. But these differences are not purely due to income differences (Monthly Labour Review 1990-<https://www.bls.gov/mlr/1990/05/art4full.pdf>). Sri Lanka’s Contribution based retirement system has currently reached its limits (Nirosha Gaminiratne (2004). Some analysis in the literature shows that the number of elderly members and pre-school

children in a household are strong predictors of the probability and financial burden of encountering out-of-pocket healthcare expenditure (Kumara and Samaratunge (2016)).

Therefore, the analysis of income and expenditure patterns of the elderly population is extremely important in developing public, health and transportation policies to accommodate senior citizens, since Sri Lanka is a country that allocates a considerable amount of revenue from the country's budget on social welfare. A deep analysis of such apportionment is seemingly absent in the current body of knowledge.

Therefore, the main purpose of this paper is to identify and provide deeper insights relevant to the income and expenditure patterns of Sri Lanka's elderly population segment and to highlight the resulting consequences on social welfare.

## **2 Research Objectives and Limitations**

### **2.1 Objective of the Study**

Main objective of the research was to identify any pattern in the income and expenses of the elderly in Sri Lanka and its impact on their key social welfare activities such as health and transportation.

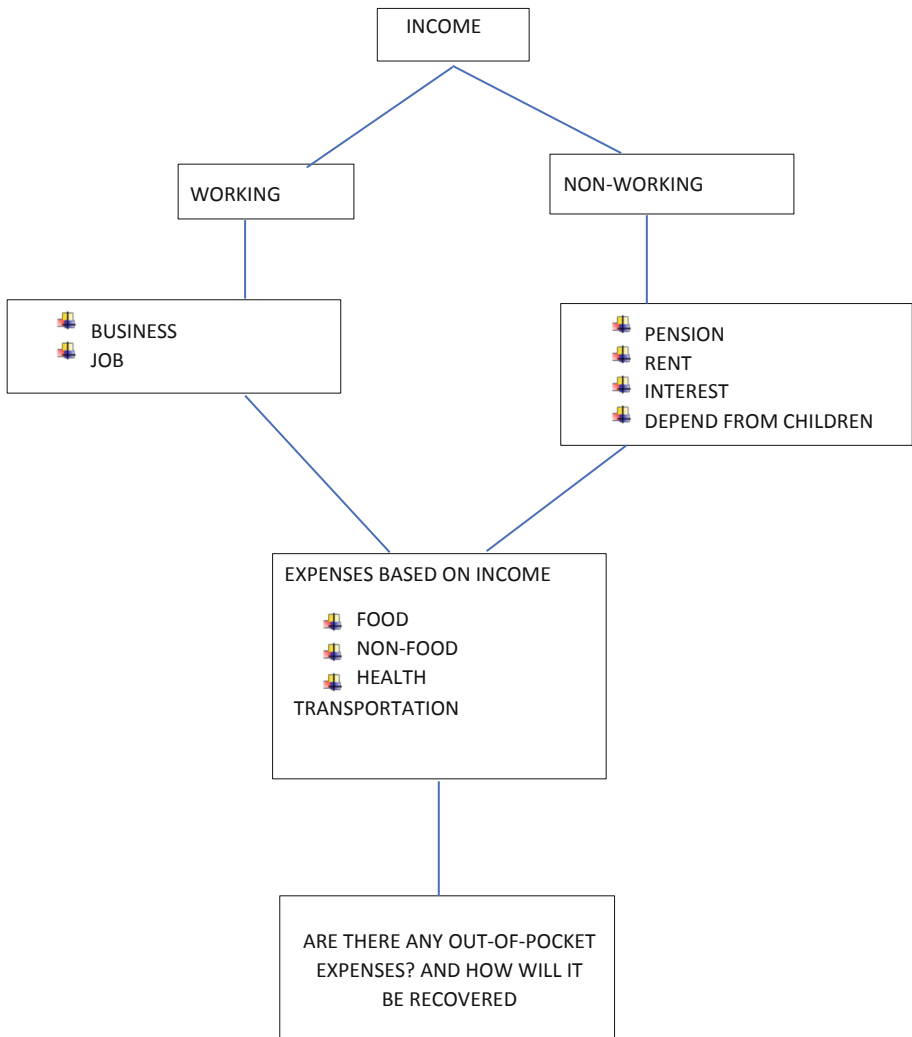
The sub objectives of the research were to find out,

1. Main income sources of elderly Sri Lankans.
2. Main expenses category of elderly Sri Lankans.
3. Identify changes in living conditions of Sri Lankan elders.
4. Identify dependencies of elderly Sri Lankans.

### **2.2 Limitations of the Study**

1. A sample of thirty elders were used in the analysis.
2. Majority of the respondents represented Colombo district located in Western Sri Lanka.
3. Time Constraint of the study.

### 3 Conceptual Framework



### 4 Methodology and Data

This research endeavor was conducted using a primary data survey in order to answer specific questions linked with the lifestyles of the Sri Lankan elderly cohort. In order to achieve the objectives of the research, a convenient sample of the elderly population was chosen as respondents. It did consist of 30 members randomly chosen from desired eligible population segment. Since the sample size was equal to 30 it meets the central limit theorem (CLT) in statistics. Therefore, based on the CLT a logical and a reasonable assumption can be made that with respect to the chosen sample the distribution of the means will be approximately normal (Hogg and Craig, 1995). Survey was conducted within a period of two weeks to avoid the impact of inflation on the results of the research. As the survey was performed during the period when COVID-19 pandemic prevailed, online questionnaires were used and mailed through the internet. Format of the survey tools can be found in the attached appendix of this paper. Link given below leads to the corresponding survey.

**Survey Link:** <https://www.surveymonkey.com/r/65JD8NM>.

The sample gathered represented 7 districts (distinctly demarcated areas in Sri Lanka) and did consist of 30 elders of which 67% represented Colombo district and the rest coming from the districts of Gampaha, Kalutara, Rathnapura, Galle, Badulla and Batticaloa. Respondents represented 7 out of 25 districts (nearly one third of the demarcated areas) in Sri Lanka. The proportions are depicted in Fig. 1 below. Since the objective of the research was to obtain a snapshot view of the whole country, the discrepancies in the income and consumption patterns among districts are not considered.

The sample represented elderly population ranging from the age 55 to 84 from which 47% is in between the age of 61–70 years, 46% is in between 55–60 and 7% being in between 71–87 years (Fig. 2).

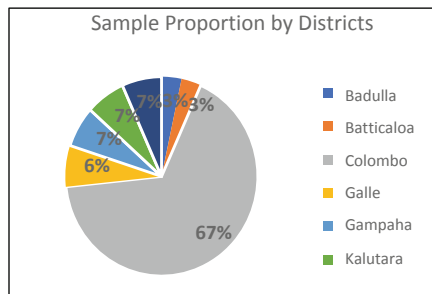
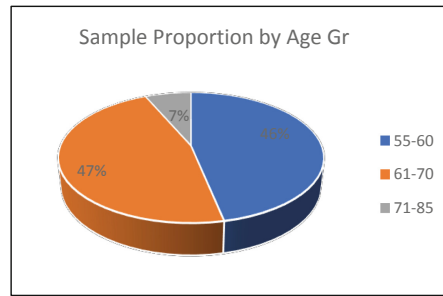


Fig. 1. .



**Fig. 2.** .

## 5 Data Analysis and Discussion

### 5.1 Income Analysis

In this study income refers to value items received either in monetary or non-monetary terms by all the elders within the chosen sample. Income earners could also be categorized into two major streams as working and non-working. But income sources of the workers consist of wages and salaries, business activities and non-working income. Non-working income comprises of interest earnings, pension payments, foreign transfer revenue, welfare payments and unforeseen income gains (e.g.: lottery wins, compensations etc.). These revenue sources are identified as most common monetary income of senior citizens in Sri Lanka.

The non-monetary income is the estimated value of goods and services received in kind and consumed. As the valuation of such goods and services can be subjective, non-monetary income has not been taken into consideration for this study and therefore the research focus is only on monetary income.

The survey reveals that the average income per month of an elder in Sri Lanka was Rs.248,833 in 2020. The median income per month of an elder in Sri Lanka has been reported at Rs.60,000 in 2020. Average income is a point-estimate, and it is calculated by dividing the total income by the number of elderly populations within the same domain. Median income is the amount that divides the sorted household income distribution into two equal groups, i.e., half having income above and other half having income below that amount.

As it can be seen in Table 1, two main contributions to mean income are reported from salary/wages and business activities with proportions of 39% and 38% respectively.

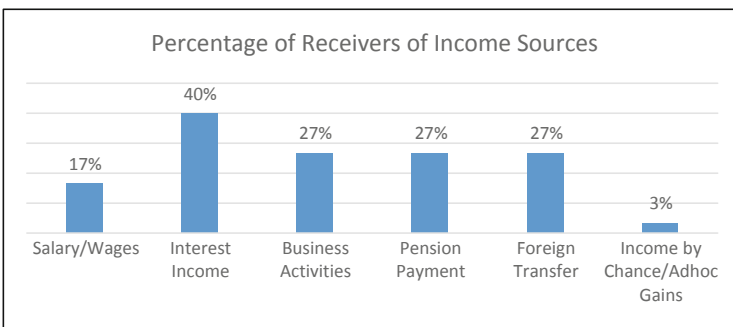
Figure 3 below shows the results of the survey with regard to percentage of receivers from different income sources. It shows that 27% of the elders solely rely on pension payments. In Sri Lanka there is a high dependency ratio. Also pension benefits are not indexed to prices and wages. Therefore, real value of income gets deteriorated at the time of retirement. Since Pension benefits are generally deemed to be insufficient to meet the needs of elderly, their consumption becomes limited.

**Table 1.** .

Source of Income	Mean	Income share
Salary/Wages	97,400	39%
Interest Income	31,600	13%
Business Activities	94,833	38%
Pension Payments	9,167	4%
Foreign Transfers	12,500	5%
Income by Chance/Adhoc Gains	3,333	1%
<b>Total Income</b>	<b>248,833</b>	<b>100%</b>

Also, it is noted that 40% have interest income as their revenue source, which is generated from their investments. These investments are commonly made from the money they saved in their younger age and their offspring’s allocation of funds for their parents to generate monthly income. The survey results further show that 10% have interest income as their only income source which is volatile with the fluctuations in interest rates. Accordingly, their monthly income is not fixed and becomes variable in the medium-term.

Receiving income from relatives and well-wishers who are residing permanently or for work purposes in a foreign country is another common form of income for Sri Lankan elders. Twenty seven percent (27%) of elderly surveyed receive income from foreign transfers out of which 13% receive it as their sole revenue. It indicates a considerable number of elders are depending on children and relatives. Furthermore, it is worthwhile to note that these remittances not only support the elders, but also the Sri Lankan economy in terms of foreign exchange.



**Fig. 3.** .

Although salary wages are the largest proportion of mean income, it is received by only 17%, and it has also been found that the employed elders are below 60. The salaries and wages become the largest portion of mean income due to very high

salaries earned by elderly people because of their working experience and senior positions held within their workplace hierarchy.

**Remark:** From Fig. 3, it is evident that the sum of all the percentage of receivers across all income sources do not add up to 100%. It adds up to 141%. Reason for the sum exceeding 100% is because some of the senior citizens receive an income from multiple sources. Therefore, the income probability distribution of individuals will comprise of both mutually exclusive as well as intersecting revenue sources. The implication of it is that certain individual senior citizens will get included in multiple income source categories contributing towards their percentage of receivers. It is evident in the boosted percentage values of senior citizens in Fig. 3.

Only a total of 44% generate income by actively participating in business and employment. These elders can be considered as the ones who are contributing to the country's GDP even at their retirement age by creating economic value to the country. The balance 57% generate income through sources that do not require their physical involvement. Given that the worsening health conditions as a human age explains the reason for majority not being reliant on physical activities to survive.

**Table 2.** .

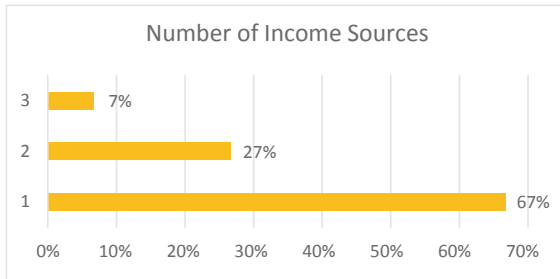
	1st Quintile	2nd Quintile	3 <sup>rd</sup> Quintile	4 <sup>th</sup> Quintile	5th Quintile
Mean Income per month (Rs.)	29,166.67	40,000.00	60,000.00	105,000.00	1,010,000.00
Share of income (%)	2.3%	3.2%	4.8%	8.4%	81.2%
Cumulative share of income (%)	2.3%	5.6%	10.4%	18.8%	100.0%
Cumulative % of Population	20%	40%	60%	80%	100%

The elderly per month income values, are arranged in ascending order and thereafter divided into five groups with equal frequencies. Such a group is defined as an income quintile. As per Table 2, out of a representative sample of 30 elders, the richest 20% generate 81.2% of the total income which indicates a very high level of income inequality among elders.

It is a phenomenon known as a Pareto probability distribution in mathematical statistics or the Pareto law in wealth distribution. Certain literature sources state it as the 80-20 rule or the power law probability distribution. The Pareto law of wealth distribution states the density of distribution of wealth  $W$  is proportional to the power of wealth  $W^{-\alpha - 1}$ , where  $\alpha$  is the Pareto index parameter (shape parameter of Pareto distribution) generally defined in practice within the range of 1.6 to 2.4 if rather low levels of income have been adjusted upwards for convenience. In the above analysis  $\alpha$  will yield approximately a value of  $\log_4 5 \approx 1.16$  precisely reflecting the 80-20 rule (Wold and Whittle, (1957) and Pareto (1898)).

It is interesting to note that as per the rules of statistical inference governed by CLT a sample of 30 elders provide an approximate pseudo-Pareto probability distribution of elderly income with a sampling error of negligible magnitude. On such a basis a larger sample would yield a much more accurate Pareto probability distribution of elderly income with respect to Sri Lanka.

The richest 20% of Sri Lankan elders in such a context generate their income from business and employment as experienced employees in senior roles in organizations and established business leaders. Poorest 20% accounts for only 2.3% of total income with an average of Rs 29,166.67. These are small scale business owners and retired pension payment receivers.



**Fig. 4. Note:** Certain percentages in Fig. 4 have been rounded up by the utilized software thereby resulting in the sum of percentages exceeding 100% by a negligible percentage value.

Figure 4 shows the proportion of elders that obtain income from a variety of sources. The maximum number of income sources is three generated by only 7%. Sixty seven percent (67%) of elders as per Fig. 4 have only one income source to fund their consumption.

**5.2 Consumption Analysis**

The survey reveals that the average consumption per month of an elder in Sri Lanka was Rs. 63,600 in 2020. The median consumption per month of an elder in Sri Lanka has been reported as Rs. 40,000 in 2020. Average consumption is a point estimate, and it is calculated by dividing the total consumption in a domain by the number of elderly population members in the same domain. Median consumption is the amount that divides the sorted household income distribution into two equal groups, i.e. one half having consumption above and the other half having consumption below that measure.

Food ratio has been computed by dividing elderly total food expenditure (excluding expenses on liquor, narcotics drugs and tobacco) by total expenditure as given below. Out of expenditure, food ratio is 33.7% being the largest proportion of expenditure as it is an essential cost component.

$$\text{Food Ratio} = \frac{\text{(Expenditure on food)}}{\text{Total expenditure}} * 100$$



**Table 3.**

Consumption	Mean	%
Food	21,433	33.7%
Housing (Rent and Water bill)	5,717	9.0%
Fuel & Light	2,667	4.2%
Clothing, Textiles & Foot wear	2,333	3.7%
Health & Personal care	3,567	5.6%
Transport & Communication	3,500	5.5%
Cultural & entertainment	1,767	2.8%
Household Services (Laundry, servant charges etc.)	9,067	14.3%
Durable household goods (Furniture, electronics etc.)	500	0.8%
Donations	117	0.2%
Liquor, Narcotic drugs & Tobacco	1,733	2.7%
Payment of debt	5,333	8.4%
Other expenses	5,867	9.2%
Total Consumption	63,600	100%

Non- food expenditure is 66.3% of total expenditure. The largest non-food expenditure is the expenditure that is spent on household services. It shows the additional needs of elders to fulfill their household requirements such as laundry and cleaning. Out of 30 elders, 93% incur expenses on health and personal care with an average of Rs. 3,567. This illustrates that the expenditure on health care has become a necessity for elders.

### 5.3 Communication and Transportation Analysis

From Table 3 it is evident that the transportation and communication ratio is 5.5%. The implication is that communication represents a larger segment of the 5.5%, since most senior citizens will spend more on their communication bills in trying to reach out to their loved ones who live separately. Furthermore, senior citizens will be either less mobile or immobile during their twilight years. Thus, the culmination of all such factors will lead towards the reasonable assumption that the elderly cohort living in Sri Lanka will be responsible for a transportation ratio of about 2–3% approximately. An apportioned percentage of such magnitude will not be sufficient to provide significant accessibility to most of the senior citizens. Therefore, further research is required for a developing country like Sri Lanka to optimize its transportation for senior citizens.

Research done in the past has indicated that the availability, accessibility, and affordability of transport play an important role in affecting older people's travel preferences and addressing their mobility needs (Lin et al. 2014, Guzman and Oviedo 2018, Wong et al. 2020). A key issue that became apparent in the provision of transport infrastructure and services is that future transport policies for addressing elderly citizens' mobility

needs require the consideration of transport desired by senior citizens than exclusively focusing on what they needed (Lin and Cui, 2021). When providing legislative and institutional support for senior citizens' mobility, it is vital to realise that the objective of transport alternatives such as transit is not to simply move elderly adults from origin to destination, but to provide a desirable transport system that is affordable, efficient, convenient, comfortable and enjoyable, and fully accommodates senior citizens' travel characteristics and concerns (Lin and Cui, 2021). Accelerating the development of such transport infrastructure and services that are financially acceptable, should be a world-wide imperative (Lin and Cui, 2021). Furthermore, it will be a feasible solution for Sri Lanka's elderly citizen cohort provided a revised transportation ratio is capable of providing the required fiscal stimulus.

## 6 Conclusion

Main income sources are salary, wages and business activities. Nearly a quarter of the elderly population receive a pension at retirement. A considerably large proportion of elderly rely only on one income source to finance their consumption. A very high-income inequality was noted among the income of elderly as around 80% of the total income is generated by the richest 20% of senior citizens.

The main expenditure as per the analysis is the food ratio component. Out of non-food expenditure a large portion is incurred on household services such as laundry and domestic help. Expenditure on health and personal care is incurred by over 90% of the elderly and has become an essential expense. There are no out of pocket expenses which means that they have an average income which is greater than the average expenditure.

According to the study it reveals more than one third of the elderly population are independent and have their own income. Hence, their consumption cost is not borne by the economy. It can be seen more than 25% of elders solely depend on pension income. According to several studies aging population will be increased in future and with the increase expenses on pension payments of the government is also expected to rise. Therefore, the Sri Lankan government must pay attention to this area when making future public policies relevant to state sector retirees.

## Annexure

Income and Consumption Survey of Elderly Population in Sri Lanka – Survey Format

(Fill this form only if you are 55 or above)

Name: .....

Age: .....

District: .....

Monthly Income (Rs.): .....

### **Income Sources (Rs.):**

Salary/Wages: .....

Interest Income: .....

Business Activities: .....

Pension Payment: .....

Samurdhi/Disability Payment: .....

Foreign Transfer: .....

Other Cash Income: .....

Income by Chance/Adhoc Gains: .....

Monthly Expenditure (Rs.): .....

### **Expenditure Type (Rs.)**

Food: .....

Housing (Rent and Water bill): .....

Fuel & Light: .....

Clothing, Textiles & Foot wear: .....

Health & Personal care: .....

Transport & Communication: .....

Cultural & entertainment: .....

Household Services (Laundry, servant charges etc.): .....

Durable household goods (Furniture, electronics etc.): .....

Donations: .....

Liquor, Narcotic drugs & Tobacco: .....

Payment of debt: .....

Other expenses: .....

## References

- Attanayake, C.: The Theory of Demographic Transition and Sri Lanka's Demographic Experience (1984)
- De Silva, W.I.: Construction and Analysis of National and District Life Tables of Sri Lanka 2000–2002 (2008)
- De Silva, W.I.: The age structure transition and the demographic dividend: an opportunity for rapid economic take-off in Sri Lanka. *Sri Lanka J. Adv. Soc. Stud.* **2**(1), 3–46 (2012)
- Gaminiratne, N.: Population ageing, elderly welfare and extending retirement cover: the case study of Sri Lanka. ESAU Working Paper 3. Overseas Development Institute, London (2004)
- Guzman, L.A., Oviedo, D.: Accessibility, affordability and equity: Assessing 'pro-poor' public transport subsidies in Bogotá. *Transp. Policy* **68**, 37–51 (2018)
- Hogg, R.V., Craig, A.T.: Introduction to mathematical statistics, 5th edn. Englewood Hills, New Jersey (1995)
- Kumara, A.S., Samaratunge, R.: Patterns and Determinants of Out-of-pocket healthcare expenditure in Sri Lanka: evidence from household surveys. *Health Policy Plan.* **31**(8), 970–983 (2016)
- Lin, D., Cui, J.: Transport and mobility needs for an ageing society from a policy perspective: review and implications. *Int. J. Environ. Res. Public Health* **18**(22), 11802 (2021)
- Lin, T.G., et al.: Spatial analysis of access to and accessibility surrounding train stations: a case study of accessibility for the elderly in Perth, Western Australia. *J. Transp. Geogr.* **39**, 111–120 (2014)
- Mendis, A.: Sri Lanka Country Report. ESID/HLM-MIPAA/Macao: ESCAP (2007)
- Menike, H.R.A.: Important features of the elderly population in Sri Lanka. *Res. Proc.* **2**(2), 29–38 (2014)
- Pareto, V.: Cours d'économie politique. *J. Polit. Econ.* **6**, 49–52 (1898). <https://doi.org/10.1086/250536>
- Roberts, P., Babinard, J.: Transport Strategy to Improve Accessibility in Developing Countries. World Bank Group, Washington, DC, USA (2004)
- Siren, A., Haustein, S.: Baby boomers' mobility patterns and preferences: what are the implications for future transport? *Transp. Policy* **29**, 136–144 (2013)
- Wold, H., Whittle, P.: A model explaining the pareto distribution of wealth. *Econometrica* **25**(4), 591–595 (1957). <https://doi.org/10.2307/1905385>
- Wong, R.C.P., Yang, L., Szeto, W.Y., Li, Y.C., Wong, S.C.: The effects of accessible taxi service and taxi fare subsidy scheme on the elderly's willingness-to-travel. *Transp. Policy* **97**, 129–136 (2020). <https://www.bls.gov/mlr/1990/05/art4full.pdf>



# Bayesian Consideration for Analyzing Employee's Motivation: Evidence from Vietnam

Bui Huy Khoi<sup>1</sup>(✉) and Nguyen Ngoc Thach<sup>2</sup>

<sup>1</sup> Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam  
buihuykhoi@iuh.edu.vn

<sup>2</sup> Banking University of Ho Chi Minh City, Ho Chi Minh City, Vietnam  
thachnn@buh.edu.vn

**Abstract.** The author implements the topic to study the factors affecting the work motivation of employees. The research sample was selected by a non-probability method. The sample size is expected to be 300 samples working in the Sapo Technology Joint Stock Company, Vietnam. We processed the data by the Bayesian statistical method to give the optimal model. From there, give specific comments and solutions to solve the limitations of the topic. The research results have shown that the factors affecting the work motivation of employees include 3 factors: salary and welfare (SW), working features (WF), and corporate culture (CC) have a probability of 100%. Working environment and conditions (WEC) and training and promotion at work (TPW) have a probability of 4.4% and 20%. The paper uses the optimum selection by Bayesian consideration for analyzing employee motivation.

**Keywords:** BIC Algorithm · Salary and benefits · Working environment and conditions · Nature of Work · Training and career advancement · Corporate culture · Achievement and recognition · Relations with colleagues and leaders

## 1 Introduction

In today's social conditions, human resources are considered the determining factor in the success or failure of an enterprise. An enterprise with dynamic and creative human resources will create useful values in the development strategy of the enterprise. In order to effectively manage human resources, the first thing that managers must consider is people as the central factor of development, creating conditions to promote the full potential of each person. When managers understand their motivation, it is the basis for managers to motivate employees, motivate employees to work, and maintain this motivation, above all, keep talented employees. In addition, in recent years, the Covid pandemic broke out all over the world and in Vietnam in particular, resulting in the workforce being severely affected.

According to Today News quoted from Anphabe labor market research, an average enterprise will lose about 51% of its talent after working time. According to the forecast,

the employee turnover rate in enterprises in 2018 is 20%, of which 19% of employees feel disengaged and decide to leave, 1% of employees are engaged but still leave because of the lack of cohesion have a better chance. This rate is believed to be the highest in recent years. Also, from today's news sources, the matter is more worrying, up to 31% of human resources, although not engaged, have no intention of leaving. This is a group of employees who go to work but lose motivation and lack effort, creating many internal challenges for both the culture and performing the business. With the average core team at the company remaining at only 49%, this is a difficult and worrying problem for today's young human resources.

Citing a report by Navigos Group – a corporation providing human resource recruitment services in Vietnam, which currently owns the online job search site VietnamWorks about the need to find a job after Covid-19. The analysis is based on the opinions of 400 businesses and 1,200 job seekers who participated in the survey in August 2021. The results showed that up to 41.5% of workers said they had quit their jobs and had no new jobs. Regarding the reason for leaving, more than 30% of the candidates said they were in the category of the company's staff reduction. Next, the reason employees quit their jobs was due to salary cuts and welfare regimes accounted for nearly 25%. More importantly, nearly 52% of workers said they would change jobs after the Covid-19 pandemic ended. Besides, more than 30% of employees decided to still work at the company if the salary and welfare regime remained the same. 11% of workers will ask for a salary increase and the welfare regime will remain the same after the epidemic ends.

Before the problems that businesses are facing, Sapo Technology Joint Stock Company also has many difficulties in maintaining the current staff. Established on August 20, 2008, with passion and desire for success and clear direction, Sapo quickly asserted its leading position in retail and e-commerce with 2 main products: The main ones are Bizweb and Sapo. In its 13-year development journey (from 2008 to 2021), Sapo always tries its best with the high goal of bringing satisfaction to customers through the most optimal products and technology solutions for sale. Row. Besides that, success, the company's human resource management is still limited, especially related to the maintenance of human resources of the company. In addition, due to the complicated development of the Covid 19 pandemic, which negatively affects employees, due to the sudden impact, the company has not done well in managing employees. Currently, there is no research on human resource management for the company. The article uses the optimum selection by Bayesian consideration for analyzing employee motivation.

## **2 Literature Review**

### **2.1 The Concept of Work Motivation (WM)**

Working motivation inspires people to work enthusiastically, helping them promote their inner potential, and overcome challenges and difficulties to complete the job in the best way (Visser 2005). Work motivation is the desire and willingness of employees to increase efforts to achieve organizational goals. Personal motivation results from many resources operating simultaneously in a person and his or her living and working environment (Purwantoro and Bagyo 2019). Work motivation is an individual's desire and willingness to promote and direct his or her efforts to achieve personal and organizational

goals (Bosset and Bourgeois 2015). Based on a synthetic model of previous domestic and foreign research results. After consulting experts, the author proposed a research model with 7 suitable factors affecting the work motivation of employees including Salary and welfare, working environment and conditions, Working features, Training and promotion at work, Corporate culture, Recognition, and Relations with colleagues and leaders.

## 2.2 Salary and Welfare (SW)

Income and welfare are expressed in basic needs in Maslow's (1943) hierarchy of needs. Wages and benefits can be seen as a return or reward that employees expect to receive from their work. Wages and salaries are considered the minimum requirements of a job. When employees receive fair pay or rewards, it makes them feel satisfied and motivated. Money is known as a universal remedy that can solve most problems (Gupta and Subramanian 2014). In addition, cash payments such as wages and salaries play a very important role in management because it has a powerful impact on employees (Gupta and Subramanian 2014).

*H1: Salary and welfare have a positive influence on employee motivation.*

## 2.3 Working Environment and Conditions (WEC)

Working environment and working conditions are factors that significantly affect the level of employee motivation. According to Maslow's (1943) hierarchy of needs theory, the level of safety and security needs appears after the satisfaction of biological and physiological needs. When these needs are met, it will motivate the Staff. Components of this need include the safety, stability, protection, and orderliness of the work environment. A positive and safe work environment can bring many benefits to a company. The work environment can be separated into two components, namely physical and psychosocial.

According to Vischer (2007), that emphasizes that the physicality of the work environment plays an essential role. When working in a favorable environment, employees will be motivated to complete their work with total energy and concentration. Conversely, employees will not perform well when they are in an unsafe and distracting work environment. Good working conditions cannot by themselves motivate employees, but they can determine employee performance and productivity (Hanke 2021). Therefore, a company should establish a suitable and comfortable working environment to improve the work efficiency of employees.

*H2: Working environment and conditions have a positive influence on employee motivation.*

## 2.4 Working features (WF)

Perry and Porter (1982) suggested that workplace characteristics are related to what to do in the workplace, which matters in employee motivation. Nel et al. (2004) agree that job characteristics such as stability or challenge, creativity or pressure... Have a significant influence on work motivation. Tang and Do (2019) have shown those job characteristics and responsibilities are factors that affect employees' work motivation.

According to Wright (2003), workers who do not consider their work to be meaningful have little reason to motivate themselves to do their job. In addition, companies can convince employees of the importance of work by giving interesting reasons for their work (Locke and Latham 1990).

Supervisors can convince employees of the importance of their work by associating their work with the goals of the organization, i.e. demonstrating their importance to them. Their work to achieve organizational goals (Wright 2001). This is especially important for public institutions (Wright 2004) because the work is community-based, helping others and serving the social interests of public officials. Based on these theories, the author proposes the following hypothesis:

*H3: Working features have a positive influence on the employee's work motivation.*

## **2.5 Training and Promotion at Work (TPW)**

Training can continuously improve employees' knowledge, skills, and attitudes, which are essential for them to do their jobs effectively and efficiently (Armstrong and Taylor 2020). Aswathappa (2000) mentioned that training can also improve employees' abilities and behavior and make them perform well. Training can be used to retain key talent and ensure the personal development of employees in a particular career. So training and development can be said as a subsystem of the organization that emphasizes improving employee performance. Employee capabilities can be improved through training (Purcell et al. 2008) Training is a way for employees to expand their knowledge and fill capacity gaps so that employees can achieve their goals. Individuals and organizations better (Mugenda and Mugenda 2008; Springer 2011). In addition, training can help an organization identify employee performance and understand what needs to be improved and where (Springer 2011). Not only can organizations better understand their employees, but they can also identify what capabilities need to be improved.

In addition, training and development can foster confidence and motivate employees. This is because it can enhance employees' understanding of how their work aligns with the organization's goals, mission, and structure. When employees understand that their mission is important to the success of the organization, they feel motivated. By participating in training and development activities, employees can reduce their fear of trying new tasks, as well as reduce their frustration, conflict, and stress levels.

Training has become a major activity of many companies because they believe that well-trained employees can improve their performance (Sultana et al. 2012). Beier and Kanfer (2009) emphasize that companies can obtain suitable training facilities to improve the skills and knowledge of employees. Inevitably, when employees are adequately prepared through training and well-executed activities, it can help a company gain a competitive edge over its competitors (McDougall and Beattie 1998). In summary, training and development are important factors affecting employee motivation and organizational success (Odukah 2016).

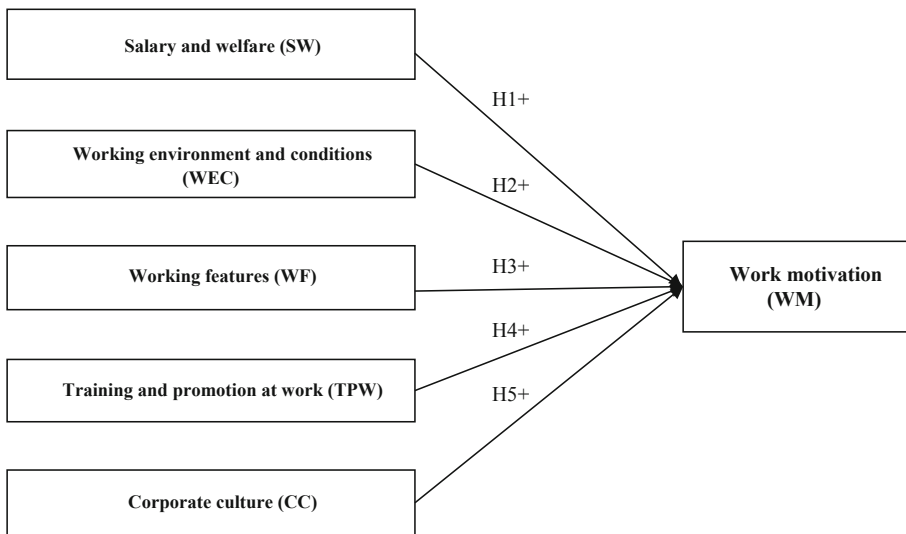
*H4: Training and promotion at work have a positive effect on employee motivation.*



## 2.6 Corporate culture (CC)

Organizational culture is an amalgamation of values, myths, heroes, and symbols that means a lot to the people who work there (Carnahan 1980). Morrison (2006) defines culture as ‘the deeper level of basic assumptions and beliefs shared by members of an organization, which operates unconsciously and which is fixed’ take for granted the organization’s view of itself and its environment”. These assumptions are learned responses to problems with a group’s existence in the external environment and problems with its internal integration. This definition sets the stage for exploring the functions of culture and the relationship between an organization’s work and its culture (Schein 2010). Durante and Gökçe (2017) show corporate culture includes a system of meanings, values, key beliefs, ways of perception, and thinking methods that are united by all members of an organization and have an influence on a wide range of perceptions and actions of each member.

*H5: Corporate culture has a positive influence on employee motivation.*



**Fig. 1.** Research model

All hypotheses and factors are shown in Fig. 1.

## 3 Method

### 3.1 Sample Size

According to Tabachnick and Fidell (2001) for the best regression analysis, it is necessary to ensure the sample size is  $N \geq 8m + 50$ . In which:  $N$ : is the sample size, and  $m$ : is the number of independent variables. Corresponding to the formula  $N > = 8m + 50$ , the number of survey samples is  $8 * 6 + 50 = 98$  samples. On that basis, to meet the

research process, the author chose a convenient sampling method to easily access the model, so the survey sample was distributed to several 295 working Sapo Technology Joint Stock Company, Vietnam. We have officially surveyed by online survey (online), starting from January 10 to June 12, 2022. Table 1 shows the sample characteristics and statistics.

**Table 1.** Statistics of Sample

Characteristics		Amount	Percent (%)
Sex and Age	Male	130	44.1
	Female	165	55.9
	18–30	170	57.6
	31–45	86	29.2
	>45	39	13.2
Education	Diploma	52	17.6
	Degree	134	45.4
	Post-graduate	92	31.2
	Other	17	5.8
Income	<5 VND mils	25	8.5
	5–7 VND mills	70	23.7
	7–10 VND mills	111	37.6
	>10 VND mills	89	30.2
Job	Business executive	135	45.8
	Team Leader	43	14.6
	Technicians	63	21.4
	Management	36	12.2
	Other	18	6.1

According to statistics, all 295 survey questionnaires were collected, and the research results showed that 130 respondents were male, accounting for 44.1% and the remaining 165 were female, accounting for 55.9%. The distribution of sample structure had a minor difference between the sex groups. Statistical results on age showed that from 18 to 30 years old had 170 votes, accounting for 57.6%, aged 31–45 years old had 86 votes, accounting for 29.2%, and those aged > 45 years old had 39 votes, accounting for 13.2%. They also show that 52 votes, accounting for 17.6%, are employees with a diploma level of education, employees with university degrees have 134 votes, accounting for 45.4%, employees with Post-graduate have 92 votes account for 31.2% and the last was qualified staff not in the above categories with 17 votes accounting for 5.8%. The results of income statistics show that 25 votes, accounting for 8.5%, are employees with income < 5 million, there are 71 votes, accounting for 23.7% of employees with income from

5 to 7 million, 111 votes, accounting for 37.67% of employees with income. Income from 7 to 10 million and finally 89 votes, accounting for 30.2% of employees with an income of > 10 million. They show that 135 votes, accounting for 45.8%, are business executives, 43 votes account for 14.6% of group Team Leaders, 63 votes account for 21.4% of technicians, 36 votes account for 12.2% management positions, and 18 votes, accounting for 6.1%, belong to other positions.

### 3.2 Bayes' Theorem

Let  $H$  be the hypothesis and  $D$  denote the actual data obtained from the collection. Bayes' theorem (Bayes 1763) states that the probability of  $H$  given  $D$  occurs, denoted as  $P(H|D)$ , is:

$$P(H|D) = \frac{P(H) * P(D|H)}{P(D)} \quad (1)$$

The probability of the hypothesis before collecting data is called  $P(H)$ .  $P(D|H)$  is the probability that the data happens under the correct hypothesis  $H$ ;  $P(D)$  is the distribution of the data in Eq. 1 (Thang 2021).

### 3.3 Bayes Inference

According to Gelman and Shalizi (2013), based on the Bayes theorem, we can see that the inference of Bayes has 3 types of information: information we want to know [posterior information], the information we already know [prior information], and practical information [likelihood]. Here, "information" can be understood as probability or distribution in Eq. 2. Therefore, Bayesian inference can be generalized:

$$\text{Posterior information} = \text{Prior information} \times \text{Likelihood} \quad (2)$$

### 3.4 Selection of the Model by the Bayesian Model Averaging

Usually, to simply define a model for a research problem, one gives only a single model (the model includes all the collected variables) to estimate and then deduce, as if that model were the model most suitable for the data. Therefore, the method can ignore other models built with some variables from the set of collected variables, and some of those models may be more suitable. Therefore, it is necessary to survey and compare the models of a research problem to find the actual most suitable model for the data (which can also be interpreted as the "best" model) (Raftery 1995). The Bayesian statistical model selection method is the Bayesian mean model method (BMA), which uses posterior probabilities and the BIC index to measure the model (Raftery 1995). The advantage of using the BMA method is the ability to take the model uncertainty into account by considering all models of the study.

### 3.5 Bayesian Information Criteria

In Bayesian statistics, prior knowledge serves as the theoretical underpinning, and the conclusions drawn from it are mixed with the data that have been seen (Thach 2020). According to the Bayesian approach, the probability is information about uncertainty; probability measures the information’s level of uncertainty (Kubsch et al. 2021). As a result, the Bayesian approach is increasingly popular, particularly in the social sciences. Bayesian statistics became a popular tool with the rapid development of data science, big data, and computer computing (Kreinovich et al. 2018). The BIC is a significant and practical metric for selecting a complete and simple model. Based on the BIC information standard, a model with a lower BIC is picked. The best model will end when the minimum BIC value is attained (Kaplan 2021).

First, the posterior probability  $P(\beta_j \neq 0|D)$  given by variable  $X_j$  with  $(j = 1, 2, \dots, p)$  indicates the possibility that the independent variable affects the occurrence of the event (or a non-zero effect).

$$P(\beta_j \neq 0|D) = \sum_{M_k \in A} P(M_k|D) * I_k(\beta_j \neq 0) \tag{3}$$

where A is a set of models selected in Occam’s Window described in Eqs. 3 and  $I_k(\beta_j \neq 0)$  is 1 when  $\beta_j$  in the model  $M_k$  and 0 if otherwise. The term  $P(M_k|D) * I_k(\beta_j \neq 0)$  in the above equation means the posterior probability of the model  $M_k$  not included  $X_j = 0$ . The rules for explaining this posterior probability are as follows (Raftery 1995): Less than 50%: evidence against impact; Between 50% and 75%: weak evidence for impact; Between 75% and 95%: positive evidence; Between 95% and 99%: strong evidence; From 99%: very strong evidence;

Second, an estimate of the Bayes score and standard error is given by the formula

$$E(\beta_j|D) = \sum_{M_k \in A} \hat{\beta}_j P(M_k|D) \tag{4}$$

$$SE(\beta_j|D) = \sqrt{\sum_{M_k \in A} \{[var(\beta_j|D, M_k) + \hat{\beta}_j^2]P(M_k|D)\} - E(\beta_j|D)^2} \tag{5}$$

with  $\hat{\beta}_j$  is the posterior mean of  $\beta_j$  in the  $M_k$  model. Inference about  $\beta_j$  is inferred from Eqs. (3); (4) and (5).

## 4 Results

### 4.1 Reliability Test

Cronbach’s Alpha test is a tool to help the author check whether the observed variables of the crucial factor are reliable or not and whether the variable is good. This test reflects whether the criteria for compatibility and concordance among dependent variables in the same major factor are closely related. The higher the coefficient of Cronbach’s Alpha ( $\alpha$ ), the higher the reliability of the factor. Cronbach’s Alpha value coefficient includes the following values: 0.8 to 1: very good scale, 0.7 to 0.8: good use scale, 0.6 and above: qualified scale. If a measure has a Corrected item-total Correlation (CITC) greater than 0.3, then that variable meets the requirements (Nunnally 1994).

**Table 2.** Reliability

Factor	$\alpha$	Item	Code	CITC
Salary and welfare (SW)	0.867	Your current salary corresponds to your work results	SW1	0.727
		The fairness of wages between the employees in your company	SW2	0.566
		There is fairness and satisfaction in your company's bonus policy	SW3	0.816
		Your salary is paid in full and on time	SW4	0.634
		The company's welfare policy is clear and helpful	SW5	0.625
		Policies on health insurance, social insurance, and unemployment insurance are fully complied with by the company	SW6	0.699
Working environment and conditions (WEC)	0.870	Professional working environment suitable for you	WEC1	0.764
		You are provided with full tools and equipment at work	WEC2	0.692
		Your workplace is clean, comfortable, and safe	WEC3	0.725
		You are satisfied with the company's activities: fun, birthday, travel, annual	WEC4	0.824
Working features (WF)	0.913	You have a good understanding of the work you are doing	WF1	0.841
		You are finding the work you are doing interesting	WF2	0.839
		The job you are doing is challenging	WF3	0.704
		You can apply many skills while working	WF4	0.818
		Your current job is suitable for your ability and forte	WF5	0.680
		Your work is properly assigned by your superiors	WF6	0.697
Training and promotion at work (TPW)	0.928	You have been fully trained by the company with the skills needed to do your job	TPW1	0.854

*(continued)*

**Table 2.** (continued)

Factor	$\alpha$	Item	Code	CITC
		The company’s job promotion policy is fair	TPW2	0.884
		The company’s work provides the right process and specific instructions for you to understand when working	TPW3	0.871
		You will have many opportunities for career advancement when applying to the company	TPW4	0.724
Corporate culture (CC)	0.732	You clearly understand the vision and mission of the company	CC1	0.773
		Your company is highly appreciated for its reputation and quality of products and services	CC2	0.792
		You like the company’s culture	CC3	0.705
		You feel the company is your second home	CC4	0.097
Work motivation (WM)	0.826	You are passionate, enthusiastic, and enthusiastic about your work	WM1	0.665
		You are ready to accept the tasks assigned by your superiors	WM2	0.637
		You will work at the company despite an attractive offer from another company	WM3	0.780

$$\alpha = \frac{k}{k - 1} \left[ 1 - \frac{\sum \sigma^2(x_i)}{\sigma_x^2} \right]$$

Table 2 shows the Cronbach’s Alpha coefficient of Corporate culture (CC), Training and promotion at work (TPW), Working features (WF), Working environment and conditions (WEC), Salary and welfare (SW) for Work motivation (WM) is all greater than 0.7. This shows that the factors are all reliable. Table 2 shows that all Corrected item-total Correlation of items is greater than 0.3. CC4 (0.097) is rejected because its CITC is lower than 0.3. That shows that the items are correlated in the factor and they contribute to the correct assessment of the concept and properties of each factor. Therefore, in testing the reliability of Cronbach’s Alpha for each scale, the author found that all the observed variables satisfy the set conditions that the Cronbach’s Alpha coefficient is greater than 0.6 and the Corrected item coefficient – Total Correlation is greater than 0.3, so all items are used for the next test step.

## 4.2 BIC Deep Learning Algorithm

There have been many algorithms created and extensively explored for detecting association rules in transaction databases. Other mining algorithms, such as incremental updating, mining of generalized and multilevel rules, mining of quantitative rules, mining of multi-dimensional rules, constraint-based rule mining, mining with multiple minimum supports, mining associations among correlated or infrequent items, and mining of temporal associations, were also presented to provide more mining capabilities (Gharib et al. 2010). Big Data Analytics and Deep Learning are two areas of data science that are receiving considerable interest. As many individuals and organizations have been gathering enormous amounts of Deep Learning algorithms for Working Motivation, Big Data has become increasingly important (Najafabadi et al. 2015). BIC (Bayesian Information Criteria) was used to select the best model by R software. In the theoretical environment, BIC has been used to choose models. BIC can be used as a regression model to estimate one or more dependent variables from one or more independent variables (Raftery et al. 1997). The BIC is an important and useful metric for determining a full and straightforward model. A model with a lower BIC is chosen based on the BIC information standard (Kaplan 2021; Raftery et al. 1997; Raftery 1995). R report shows every step of searching for the optimal model. BIC selects the best 4 models as Table 3.

**Table 3.** BIC model selection

WM	Probability (%)	SD	model 1	model 2	model 3
Intercept	100.0	0.125760	-0.3417	-0.4182	-0.3416
SW	100.0	0.041293	0.5363	0.5503	0.5361
WEC	4.4	0.007844			0.0003
WF	100.0	0.034272	0.3926	0.4.265	0.3925
TPW	20.0	0.025243		-0.0529	
CC	100.0	0.033351	0.1367	0.1.589	0.1366

There are five independent and one dependent variable in the models in Table 3. Salary and welfare (SW), working features (WF), and corporate culture (CC) have a probability of 100%. Working environment and conditions (WEC) and training and promotion at work (TPW) have a probability of 4.4% and 20%.

## 4.3 Model Evaluation

According to the results from Table 4, BIC shows model 1 is the optimal selection because BIC (-5.332) is minimum. Salary and welfare (SW), Working features (WF), and Corporate culture (CC) impact Work motivation (WM) is 84.5% ( $R^2 = 0.845$ ) in Table 4. BIC finds model 1 is the optimal choice and five variables have a probability of

75.6% (post prob = 0.756). The above analysis shows the regression equation below is statistically significant.

$$WM = -0.3417 + 0.5363SW + 0.3926WF + 0.1367CC$$

Coded: Work motivation (WM), salary and welfare (SW), working features (WF), corporate culture (CC).

**Table 4.** Model Test

Model	nVar	R <sup>2</sup>	BIC	post prob
model 1	3	0.845	-5.332	0.756
model 2	4	0.847	-5.305	0.200
model 3	4	0.845	-5.275	0.044

**BIC = -2 \* LL + log (N) \* k**

## 5 Conclusions

This study uses the optimal choice of the BIC Deep Learning Algorithm for Work motivation (WM) of laborers working at Sapo Technology Joint Stock Company, Vietnam. According to statistics from the survey sample, 61.2% are male and the remaining 38.8% are female. With these 295 suitable survey samples, the author went into reliability testing and linear regression analysis. From the regression results, the author believes that perceived quality is the factor that has the strongest and most positive impact on Salary and welfare (SW) with the index  $\beta = 0.5363$  followed by Working features (WF) with  $\beta = 0.3926$ , and finally Corporate culture (CC) with a low  $\beta$  index 0.1367.

### 5.1 Implications with Salary and Welfare (SW)

According to the data of the observed variable SW1 “Your current salary corresponds to your work results” has an average value of 3.6441, so the salary affects policies and benefits for employees to attract employees to stay and contribute to the company. Policies on salary corresponding to the level of work bring trust to the business and employees and benefits for employees such as improved work results are equivalent to an increased salary, high, bring high satisfaction, and promote advancement in work. Welfare also plays a role in retaining potential candidates and also helps all employees of the company to be kept assured of striving at work.

Observable variable SW2 “Wage fairness among employees in your company” has an average value of 3.6576. Fairness among employees in the company represents competition and other factors, other incentives. If employees want to get a higher salary, it means that they have to push themselves to develop more at work. The fairness of salary



for each position is also a measure to evaluate employees at the department to contribute to the promotion, promotion, and recognition of the results of one's efforts.

With the observation SW3 "There is fairness and adequacy between your company's bonus policy" with an average value of 3.7119, fairness builds trust for each case. Fairness and transparency are two essential conditions for creating trust in a large collective. When everything is transparent and fair, there will be less personal animosity or unfairness toward employees.

With the observed variable SW4, "Your salary is paid in full and on time" with an average value of 3.8305, this is the core factor to evaluate the company that employees are working for. With a mean value of 3.8305, workers' concern about their earnings being paid on time creates professionalism and interest in individual workers. Because each worker's focus on work is salary, and when that money is paid correctly, they can use it for personal purposes and daily living.

With the observation variable SW5, "The company's welfare policy is clear and helpful." With an average value of 3.8305, benefits improve the material and spiritual life of employees, help them self-motivate and improve labor productivity. Compensation policies also help increase the reputation of the business in the marketplace, and the positive feedback of employees, and when those policies are of little use, they will preserve a high-quality workforce. Reducing burdens for employees such as social insurance, health insurance, and unemployment insurance.

With the observed variable SW6 "Policies on health insurance, social insurance, and unemployment insurance are fully complied with by the company" with an average value of 3.8068, the usefulness of these policies helps to reduce the anxiety of employees. Working with factors such as labor health, medical service costs when sick and fees paid later. These policies create peace of mind for employees, helping them to reduce some of their anxiety and fully focus on their work.

## 5.2 Implications for Working Features (WF)

According to WF1 observed variable data, "You have a good understanding of the work you are doing" has an average value of 3.7898, about understanding what you are doing is like managing your time. Each individual is unique. Knowing the work arrangements in a certain order will help the employees themselves always complete the assigned tasks, avoiding pressure to put pressure on themselves.

The WF2 observation variable "You feel the work you are doing is interesting" has an average value of 3.8542, interesting at work is a big part of the motivation for employees to stay at the company to continue. Customary. Everyone's work will be boring if they don't love it, the interest creates curiosity which becomes the motivation for employees.

The observed variable WF3 "The work that you are doing is challenging and difficult" has an average value of 4.3119 when controlling for the stability and speed of the assigned workloads, focusing on solving problems. Solving these problems will help you feel more comfortable. More challenges and difficulties help employees realize that they have to try harder for the job, helping to increase their own experience over time, the salary is also better paid, and the difficulty is more difficult. Those towels were just the initial springboards that made me stop and flinch.

The observed variable WF4 “You can apply many unique skills while working” has an average value of 3.8475, applying unique skills when working makes the job faster and simpler. Each employee will have personal strengths in a certain area, as long as they know how to fit those pieces into their current job and proactively address them. Those skills will be continuously cultivated and enhanced, making the worker a more proactive person, and simpler in solving various problems.

The observed variable WF5 “Your current job is suitable for your capabilities and strengths” has an average value of 3.9898. Each individual’s strengths will be different, but each person will have factors to evaluate their potential through the working process. The types we often meet are integration, hard work, and carefulness,... These are the factors that help your work continue evenly on gears in an enormous machine. When an employee’s abilities and personal strengths are suitable for the job, that job will be moved in a more positive direction.

The observed variable WF6 “Your work is properly assigned by your superiors” has an average value of 3.9831, the right person has the right job with full knowledge and skills to solve the job. In addition, the consistency in work results and time compliance of each individual is different. The work is assigned to the right employee with the elements to help that job achieve excellent results.

### **5.3 Implications for Corporate Culture (CC)**

According to the observed variable CC1 “You clearly understand the vision and mission of the company” has an average value of 4.6746, when working for a company, understanding the vision and mission is important to help Let employees understand who they will do, for what benefit and why they have to do so that is the mission. A vision will help an individual know how to move towards goals and achieve them. Employees who understand the two factors above will create a foundation for themselves when they are new and the company, locate the goals and requirements of the organization, and work towards a common goal.

The observed variable 4.6881 “Your company is highly appreciated for its reputation and quality of products and services” has an average value of 4.69, creating personal trust for the company. When an individual is recognized and evaluated by others, the results bring about prestige and quality. It will help that individual have more confidence in the work they are doing, creating greater motivation when mentioned. Those motivations and beliefs promote high moral values and increase work progress.

Observational variable 4.64753 “You like the company’s culture” has an average value of 4.65, corporate culture has the fourth impact on employee engagement and thorough evaluation. Employees also think that they have not been respected by the bank, so in the future, managers should have the policy to change and adjust the culture toward employees’ values so that employees can feel better and avoid negative thoughts. Deviant leave. In addition, managers should also review current cases of not being treated fairly and fairly to take corrective measures or handle them according to regulations so that employees feel they are treated fairly like other employees. Another member.

## 6 Limitations

Due to the complicated development of the Covid-19 epidemic, the author had difficulty reaching the survey subjects. The number of surveyors is limited to 295 samples; the paper-only surveys by online survey via a google form. From the above limitations, the author has proposed many further research directions such as: Because the overall representativeness of the sample is not high, it is necessary to increase the number of survey samples to ensure the generalizability of this topic. Second, it is advisable to add other factors to the topic to analyze the level of impact on employee motivation. Finally, the survey should be conducted in person rather than online. It will help the author reach the right audience and can help that audience type the right answer for that person.

## References

- Armstrong, M., Taylor, S.: *Armstrong's Handbook of Human Resource Management Practice*. Kogan Page Publishers (2020)
- Aswathappa, K.: *Human Resource and Personnel Management*. Tata Mcgraw, New Delhi (2000)
- Bayes, T.: LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, pp. 370–418 (1763)
- Beier, M.E., Kanfer, R.: *Motivation in training and development: a phase perspective*. In: *Learning, Training, and Development in Organizations*. Routledge (2009)
- Bosset, I., Bourgeois, E.: *Motivation to transfer: linking perceived organizational support to training to personal goals*. In: Gorges, J., Gegenfurtner, A., Kuper, H. (eds.) *Motivationsforschung im Weiterbildungskontext*, pp. 169–199. Springer Fachmedien Wiesbaden, Wiesbaden (2015). [https://doi.org/10.1007/978-3-658-06616-1\\_10](https://doi.org/10.1007/978-3-658-06616-1_10)
- Carnahan, P.J.: *Organizational Structure, Work Values and Conflict*. Iowa State University (1980)
- Durante, R., Gökçe, A.T.: *Organizational culture and ethics: the influence organizational and personal values have on perceptions of misconduct and the factors of whistleblowing*. In: *Encyclopedia of Strategic Leadership and Management*. IGI Global (2017)
- Gelman, A., Shalizi, C.R.: *Philosophy and the practice of Bayesian statistics*. *Br. J. Math. Stat. Psychol.* **66**, 8–38 (2013)
- Gharib, T.F., Nassar, H., Taha, M., Abraham, A.: *An efficient algorithm for incremental mining of temporal association rules*. *Data Knowl. Eng.* **69**, 800–815 (2010)
- Gupta, B., Subramanian, J.: *Factors affecting motivation among employees in consultancy companies*. *Int. J. Eng. Sci. Invention* **3**, 59–66 (2014)
- Hanke, D.: *Can employees motivate themselves? The link between peer motivating language and employee outcomes*. *The Int. Trade J.* **35**, 19–39 (2021)
- Kaplan, D.: *On the quantification of model uncertainty: a bayesian perspective*. *Psychometrika* **86**(1), 215–238 (2021). <https://doi.org/10.1007/s11336-021-09754-5>
- Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.): *ECONVN 2019. SCI*, vol. 809. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-04200-4>
- Kubsch, M., Stamer, I., Steiner, M., Neumann, K., Parchmann, I.: *Beyond p-values: using bayesian data analysis in science education research*. *Pract. Assess. Res. Eval.* **26**, 4 (2021)
- Locke, E.A., Latham, G.P.: *A Theory of Goal Setting & Task Performance*. Prentice-Hall, Inc. (1990)
- Maslow, A.H.: *A Theory of Human Motivation*. York University, Toronto, Ontario (1943)
- McDougall, M., Beattie, R.S.: *The missing link? Understanding the relationship between individual and organisational learning*. *Int. J. Train. Dev.* **2**, 288–299 (1998)

- Morrison, J.: *The International Business Environment*. Palgrave Macmillan (2006)
- Mugenda, A.G., Mugenda, A.: *Social Science Research: Theory and Principles*. Applied, Nairobi (2008)
- Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: Deep learning applications and challenges in big data analytics. *J. Big Data* **2**(1), 1–21 (2015)
- Nel, P., Van Dyk, P., Haasbroek, G., Schultz, H., Sono, T., Werner, A.: *Human Resources Management*. Oxford University Press, Cape Town (2004)
- Nunnally, J.C.: *Psychometric theory* 3E. Tata McGraw-hill education (1994)
- Odukah, M.E.: Factors influencing staff motivation among employees: a Case study of equator bottlers (Coca Cola) Kenya. *J. Hum. Res. Sustain. Stud.* **4**, 68–79 (2016)
- Perry, J.L., Porter, L.W.: Factors affecting the context for motivation in public organizations. *Acad. Manag. Rev.* **7**, 89–98 (1982)
- Purcell, Kinnie, N., Swart, J., Rayton, B., Hutchinson, S.: *People Management and Performance*. Routledge (2008)
- Purwanto, H., Bagyo, Y.: Citizenship organizational behavior ability to increase the effect of organizational climate, work motivation, and organizational justice on employee performance. *MEC-J (Manag. Econ. J.)* **3**, 195–218 (2019)
- Raftery, A.E.: Bayesian model selection in social research. *Soc. Methodol.* **25**, 111–163 (1995)
- Raftery, A.E., Madigan, D., Hoeting, J.A.: Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**, 179–191 (1997)
- Schein, E.H.: *Organizational Culture and Leadership*. John Wiley & Sons (2010)
- Springer, G.J.: A study of job motivation, satisfaction, and performance among bank employees. *J. Glob. Bus. Issues* **5**, 29 (2011)
- Sultana, A., Irum, S., Ahmed, K., Mehmood, N.: Impact of training on employee performance: a study of telecommunication sector in Pakistan. *Interdisc. J. Contemp. Res. Bus.* **4**, 646–661 (2012)
- Tabachnick, B., Fidell, L.: *Using Multivariate Statistics*, 4th edn., pp. 139–179. HarperCollins, New York (2001)
- Tang, D.S., Do, D.T.: The impact of work characteristics on bank employees' motivation in hanoi: application of job characteristics' theory of Hackman and Oldham (1980). *European J. Busin. Manage* **11**, 27 (2019)
- Thach, N.N.: How to explain when the ES is lower than one? A Bayesian nonlinear mixed-effects approach. *J. Risk Financ. Manag.* **13**, 21 (2020)
- Thang, L.D.: The Bayesian statistical application research analyzes the willingness to join in area yield index coffee insurance of farmers in Dak Lak province. University of Economics Ho Chi Minh City (2021)
- Vischer, J.C.: The effects of the physical environment on job performance: towards a theoretical model of workspace stress. *Stress Health: J. Int. Soc. Inv. Stress* **23**, 175–184 (2007)
- Visser, J.: Working with children and young people with social, emotional and behavioural difficulties: What makes what works, work. In: *Handbook of Emotional & Behavioural Difficulties*. Sage Publications, London (2005)
- Wright, B.E.: Public-sector work motivation: a review of the current literature and a revised conceptual model. *J. Public Adm. Res. Theor.* **11**, 559–586 (2001)
- Wright, B.E.: Toward understanding task, mission and public service motivation: a conceptual and empirical synthesis of goal theory and public service motivation. In: *7th National Public Management Research Conference*, Citeseer, pp. 9–11. Georgetown Public Policy Institute, Washington, DC (2003)
- Wright, B.E.: The role of work context in work motivation: a public sector application of goal and social cognitive theories. *J. Public Adm. Res. Theor* **14**, 59–78 (2004)



# The Roles of Grassroots Government and Associations Versus Internet Access in Households' Income in Vietnam

Chon Van Le<sup>1(✉)</sup> and Thuong Thi Vu<sup>2</sup>

<sup>1</sup> International University, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

lvchon@hcmiu.edu.vn

<sup>2</sup> University of Danang, Campus in Kontum, Kontum, Vietnam

vtthuong@kontum.udn.vn

**Abstract.** This paper investigates the roles played by government and associations at grassroots level and internet access in households' welfare in Vietnam. We employ the random walk Metropolis-Hastings Markov chain Monte Carlo method with data from the Vietnam Household Living Standards Survey in 2018. It seems that internet usage has made a powerful impact on people's well-being, increasing a typical household's income per capita by 34.1%. In contrast, it does no good to participate in local associations which even marginally reduce their members' income. Unfair communes that do not give subsidies to poor households as mandated by national policies harm not only poor but also non-poor families living in these communes. Local governments should be enforced to follow national laws and policies, and to help the poor.

## 1 Introduction

It has been widely accepted that local government and associations have exerted a significant influence on economic growth and poverty reduction in developing countries. Local governance is expected to improve the functioning of markets; public policies help address market failures and facilitate social transformation (Khan, 2007). The performance of local government is effectively monitored and enhanced by local associations. Greater cooperation of households in a commune tends to increase the effectiveness of not only publicly provided services but also common property management (Narayan and Pritchett, 1999). In addition, both local government and associations speed up the diffusion of information, about the availability and the proper use of seeds, fertilizer, and chemicals in farming practices, as well as of important innovations in non-farming activities.

Many indicators and sub-indicators have been developed and revised from different data sources to measure the quality of governance (Kaufmann, Kraay, and Zoido-Lobaton, 2002; Kaufmann, Kraay, and Mastruzzi, 2009). In Vietnam, two well-known aggregate measures are Vietnam Provincial Governance and

Public Administration Performance Index (PAPI) and Provincial Competitiveness Index (PCI). The PAPI surveys have been conducted annually since 2011 by the United Nations Development Program, the Vietnam Fatherland Front, and the Center for Community Support Development Studies to document citizens' assessment of governance and public administration in their localities. Meanwhile, PCI has been a yearly business survey since 2005 in a joint collaboration between the Vietnam Chamber of Commerce and Industry and the U.S. Agency for International Development to evaluate and rank the performance, capacity and willingness of provincial governments in creating a favorable environment for private sector development.

A large number of empirical researches have used these two indicators to examine the poverty-reducing effect of the quality of local government in Vietnam (Tran et al., 2019; Nguyen et al., 2021). However, both PAPI and PCI are measures of governance at the province level. They fail to account for heterogeneity at lower-level units such as districts or communes, which are believed to have an immediate impact on households' livelihood and income. This is especially true in Vietnam where traditionally 'the monarch's law loses to a village's norms' [*phép vua thua l? làng*' (in Vietnamese)].

Moreover, dissemination of information has become faster with the proliferation of the internet. The adoption of information and communications technologies (ICT) allows households to reduce transaction costs, increase market participation (Tadesse and Bahigwa, 2015), apply new technologies (Fu and Akter, 2016), and promote farming efficiency and productivity (Ogotu, Okello, and Otieno, 2014). Internet has served as a substitute for local associations and, to some extent, for local government in fostering economic development in the new era.

This paper is to investigate how government and associations at grassroots level and internet use/access affect households' income in Vietnam. Our data are from the Vietnam Household Living Standards Survey (VHLSS) conducted in 2018 by the General Statistics Office (GSO) of Vietnam. It was a nationwide sample with 46,995 households in 3,133 communes/wards which were representative at national, regional, urban, rural and provincial levels. We find that internet has a strikingly important impact than local associations on households' economic well-being. Having access to internet increases a typical household's income per capita by 34.1% while being a member of local associations causes a marginal loss of 0.2%, holding other things constant. Additionally, unfair commune governments which do not give subsidies to poor households as mandated by national policies tend to reduce welfare of people living in those communes by 2.2%.

The paper is structured as follows. Section 2 gives a brief overview of the literature on the influences of local governance, associations, and internet on households' income. Section 3 shows the econometric model. Section 4 presents the data set, descriptive statistics, and empirical findings. Conclusions follow in Sect. 5.

## 2 Literature Review

Governance is generally considered one of the critical factors determining socio-economic performance across developing countries. The Organization for Economic Cooperation and Development (OECD) (2006) affirmed that ‘good public governance helps to strengthen democracy and human rights, promote economic prosperity and social cohesion, reduce poverty, enhance environmental protection and the sustainable use of natural resources, and deepen confidence in government and public administration.’

Khan (2007) claimed that there are two different economic approaches to governance, namely, ‘market-enhancing’ and ‘growth-enhancing’ governance. The former postulates that if governments can sustain efficient markets, especially by enforcing stable property rights, a good rule of law, curbing corruption, to minimize unproductive rent-seeking activities and the crowding out of productive ones, then private sector will foster economic development. This approach is advocated by institutionalists Krueger (1974) and North (1995) and international development and financial agencies. The ‘growth-enhancing’ governance argued that markets in developing countries are intrinsically inefficient. Even the strongest political commitment cannot push underdeveloped markets to stimulate efficient resource allocation. Successful development in developing countries therefore requires competent governance of states to accelerate the transfer of assets and resources to more productive sectors and to endorse the absorption of new technologies. They would ensure productivity growth in both the private and public sectors. The East Asian Miracle, which is attributed to a large extent to these governments’ industrial policies, has been used as an evidence in support of this argument (World Bank, 1993).

The relationship between governance, public administration and economic growth has attracted a lot of attention in researches on developing countries, though they vary significantly in scope and focus. Most large-scale studies use national-level data and concentrate on different aspects of governance. Democratic institutions are more susceptible to the demands of the poor, which leads to expansion of their access to education and decreasing income inequality, but at the expense of physical capital accumulation (Tavares and Wacziarg, 2001). Better informed citizens are more likely to vote and monitor governments’ policies, public services, and administration, making public officials more accountable (Lassen, 2005). Clear tax policies and transparent legal frameworks make economies and markets perform more efficiently (Stiglitz, 2002). Since corruption lowers private domestic and foreign direct investment, increases public investment but reduces its productivity, causes more damage to new and small firms than to larger ones, creates incentives for talented persons to engage in rent-seeking activities, and raises poverty and income inequality, good governance that reduces corruption boosts economic growth (Mauro, 1995; Tanzi and Davoodi, 2000; Gupta et al., 2002).

An element of governance that has formed a fundamental building block of development and national cohesion is civil society. It fills the space untouched by the public and private sectors. Civil society includes organizations that are not

associated with government. Small agricultural producers in developing countries face considerable challenges due to changed procurement systems in which supermarkets have been increasingly dominating and to new quality and safety standards set by developed countries. Stringent requirements such as Global GAP and larger supply volumes ordered by supermarket chains have limited their participation. In addition, as developing countries have signed more free trade agreements, smallholder farmers are compelled to compete not only with their local peers and firms, but also with farmers and agribusinesses from other countries. Joining farmer cooperatives or producer organizations would enable smallholders to gain necessary market information and access to new technologies, and to enter high-value markets (Markelova et al., 2009). Local associations also defend citizen rights and interests, monitor the performance of government in its provision of public services, and stimulate effective building and management of common property. Moreover, they offer informal insurance that protects poor households from weather shocks, and encourages them to adopt innovations that are high-return but often considered high-risk (Narayan and Pritchett, 1999).

Many studies have tried to measure the quality of overall governance by aggregate indices based on a large number of sub-indicators. Campos and Nugent (1999) identified four critical institutional components: (1) the executive, (2) civil society, (3) the bureaucracy, and (4) the rule of law. In the Worldwide Governance Indicators research project, Kaufmann, Kraay, and Mastruzzi (2009) evaluated six dimensions of governance: “Voice and Accountability,” “Political Stability and Absence of Violence/Terrorism,” “Government Effectiveness,” “Regulatory Quality,” “Rule of Law,” and “Control of Corruption.” To develop these indicators, they used more than 400 individual variables from 35 separate data sources constructed by 33 different organizations throughout the world. Using this cross-country data set, Rodrik et al. (2004) suggested a positive impact of good institutional quality on economic growth. Campos and Nugent (1999) found that the prominent institution improving economic growth is the quality of bureaucracy for East Asia, but the effectiveness of rule of law for Latin America.

The governance diversity exists across countries and within countries. Several studies have examined the possible effects of governance on households’ well-being at the provincial level. United Nations Development Programme (2011) implied that there is a positive association between PAPI and Human Development Index (HDI) in Vietnam. Tran et al. (2019) showed that good provincial governance does not on average affect household per capita income, but brings greater benefits for richer households. Nguyen et al. (2021) found an opposite result where the very poor gain the most from good governance and public administration as it boosts income growth and reduces inequality. However, there have been no studies so far that consider the quality of governance at lower-level units such as districts or communes. It is believed that their behavior is heterogeneous and has an immediate impact on households’ livelihood and income.



An alternative and increasingly powerful tool that is able to provide timely, relevant, and workable information to its users at dramatically lower cost than any traditional service is the information and communications technology (ICT). Internet-using farmers can apply new farming practices (Fu and Akter, 2016), and are more likely to switch to a pesticide that is more efficacious against pests and less harmful to humans (Cole and Fernando, 2012). They can save time and costs required to verify price information from multiple sources, thus reducing price dispersion across markets and seasons. Better knowledge of products and prices that prevail in markets enables farmers to make good choice of what crops, when, how to grow, and where to sell them profitably (Jensen, 2007). Farmers participate more in markets (Tadesse and Bahiigwa, 2015) and enhance their farming efficiency and productivity (Ogotu, Okello, and Otieno, 2014). ICT also encourages rural laborers to engage in off-farm employment, diversifying their income sources. This is particularly beneficial to rural households in developing countries because it diminishes income volatility due to external shocks in agricultural production (Leng et al., 2020). Furthermore, internet has simplified e-banking, money transfers, and payment processing that offer access to financial products and services to previously financially excluded people (Lenka and Barik, 2018). Therefore, internet has served as a substitute for local associations and, to some extent, for local government in promoting economic development in the new era.

### 3 Estimation Method

So far no indices have been built to measure the quality of governance at grassroots level. A comprehensive evaluation requires a lot of data on many criteria that are not available. Though, Munda (2017) claimed that fairness in the policy process is very important because it accounts for a majority of social values, interests and desires, perspectives, distributional issues in a coherent and transparent manner. For communal governments, fairness involves more in policy implementation than in formulation. In this paper, we consider a commune 'unfair' if it fails to give subsidies to households residing in that commune who are officially labeled 'poor', thus are eligible to receive grants under the national policies.

Apart from grassroots-level government fairness, membership in local associations and access to internet, household per capita income is supposed to depend on demographic characteristics of household heads, comprising gender, ethnicity, educational attainment, age, and marital status. Other determinants are characteristics of the household such as urban/rural residence, poverty status, household size, cultivated land use rights<sup>1</sup>, the number of working adults who are self-employed or are working in farming and non-farming sectors, access to credit, and amount of subsidies received. The regression model is

---

<sup>1</sup> In Vietnam, land (including agricultural land) is owned by the state. Organizations and individuals only hold and acquire rights to use land.

$$\begin{aligned}
\ln \quad (\text{Income per capita}) = & \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Ethnic minority} + \beta_3 \text{Educ} + \beta_4 \text{Age} \\
& + \beta_5 \text{Age}^2 + \beta_6 \text{Marital status} + \beta_7 \text{Urban} + \beta_8 \text{Poverty} + \beta_9 \text{Hhsize} \\
& + \beta_{10} \text{Land area} + \beta_{11} \text{Farmers} + \beta_{12} \text{Non-farmers} + \beta_{13} \text{Self-employed} \\
& + \beta_{14} \text{Credit} + \beta_{15} \text{Subsidy} + \beta_{16} \text{Associations} + \beta_{17} \text{Internet} \\
& + \beta_{18} \text{Unfair commune} + \delta \text{Region dummies} + \varepsilon,
\end{aligned} \tag{1}$$

where *Male*, *Ethnic minority*, *Marital status*, *Urban*, *Poverty*, *Credit*, *Associations*, *Internet*, *Unfair commune* are dummy variables. *Male* is 1 if the household head is male, *Ethnic minority* is 1 if he/she is not Vietnamese or Chinese, *Marital status* is 1 if the household head is currently living with his/her spouse, *Urban* is 1 if the household resides in urban area, *Poverty* is 1 if the household was labeled ‘poor’ by its commune in the previous year, *Credit* is 1 if the household gets a loan from a formal financial institution, *Associations* is 1 if the household head is a member of local associations, *Internet* is 1 if the household head uses internet. *Educ* is the number of the head’s schooling years. A quadratic term *Age*<sup>2</sup> is added to represent the typical pattern of increasing then decreasing income over the life. *Hhsize* is the number of members in the household. *Land area* is the area of cultivated land that the household has the right to use. *Farmers*, *Non-farmers*, and *Self-employed* are the numbers of adults who work in farming sectors, non-farming sectors, and are self-employed, respectively. To take into account heterogeneity at provincial and regional levels in geography, culture, social norms, governance, etc., households are grouped into seven regions, namely, Northern Uplands, Red River Delta, North Coast, Central Coast, Central Highlands, South East, and Mekong Delta. Six regional dummies are included in Eq. (1).

## 4 Data Description and Empirical Results

Our data are compiled from the Vietnam Household Living Standards Survey (VHLSS) which was conducted nationwide in 2018 by the General Statistics Office (GSO) of Vietnam. It consists of 46,995 households in 3,133 communes/wards which are representative at national, regional, provincial, urban, and rural levels. Information was collected during four periods in four quarters (one period per quarter) through face-to-face interviews with household heads, members and key commune officials. Data on households and individuals cover demography, education, health, employment and income, housing, fixed assets and durable goods, and participation of households in poverty alleviation programs<sup>2</sup>. Due to missing data, a sample of 43,093 households is ready for analysis.

Table 1 shows that a typical household whose head is a member of local associations has an average income per capita of VND 82.68 million, lower than the one which is not (VND 98.02 million). But access to internet signifies a remarkable difference in people’s welfare. Residents in a household that uses internet enjoy an average income per capita of VND 135.42 million, more than double VND 63.73 million in a household that does not. This stereotype holds

<sup>2</sup> A sub-sample of 9,399 households were asked about consumption expenditures.

**Table 1.** Average Annual Income Per Capita by Association Membership and Internet Access

			Internet access	
			No	Yes
			VND 63.73 mil	VND 135.42 mil
Association	No	VND 98.02 mil	VND 67.27 mil	VND 143.21 mil
Membership	Yes	VND 82.68 mil	VND 59.48 mil	VND 124.06 mil

Notes: On Dec 28, 2018, \$1 = VND 23,180 or VND 1 mil = \$43.14.

consistently when association membership and internet access are considered jointly. The cross-tabulation in Table 1 indicates that income per capita is the lowest in households which join local associations but do not use internet, and is the highest in those which are non-members and use internet.

Obviously, internet usage cannot take the whole credit for income increase. It is highly correlated with other important determinants of household earnings. Panel a in Table 2 implies that household heads that use internet are nearly ten years younger and more educated than those who do not. It makes sense since young people are more willing to embrace new technology and better educated people are able to absorb new knowledge more rapidly, thus having higher demand for internet usage. In Panel b, households living in urban areas have easier access to internet than their rural counterparts. Probably economies of scale due to big customer base in towns makes internet service to be supplied faster and more conveniently to urban residents. Their proportion of internet usage is 56.99%, almost double that among rural dwellers. Since ethnic minority groups are generally less educated and live in remote areas where basic infrastructure is not well developed, their internet connectivity is rather limited. Therefore,

**Table 2.** Internet-Related Characteristics

Panel a	Internet access	
	No	Yes
Household head's		
Average age	56.6	46.7
Average number of schooling years	6.4	10.1
Panel b	% Households using internet	
Urban areas	56.99	
Rural areas	30.65	
Vietnamese or Chinese	42.59	
Ethnic minority	18.92	
Non-poverty status in 2017	41.82	
Poverty status in 2017	9.14	

**Table 3.** Descriptive Statistics of the Sample

Variable	Mean	Std Dev	Min	Max
Income per capita (VND mil)	91.33	173.90	1.89	18705.34
Male	0.74	0.44	0	1
Ethnic minority	0.17	0.38	0	1
Number of schooling years	7.83	4.19	0	22
Age	52.79	13.71	13	113
Marital Status	0.79	0.41	0	1
Urban areas	0.30	0.46	0	1
Poverty status in 2017	0.10	0.30	0	1
Household size	3.72	1.63	1	17
Cultivated land area (thousand m <sup>2</sup> )	5.79	24.77	0	2824
Number of farming workers	1.20	1.24	0	11
Number of non-farming workers	1.08	1.02	0	7
Number of self-employed	0.48	0.80	0	8
Formal credit	0.17	0.38	0	1
Subsidies for poor households (VND mil)	5.07	9.38	0	161.19
Subsidies for non-poor households (VND mil)	0.59	5.08	0	602.82
Local association membership	0.44	0.50	0	1
Internet access	0.38	0.49	0	1
Unfair commune	0.05	0.23	0	1

the divide in internet connection between the Vietnamese or Chinese (42.59%) and ethnic minority groups (18.92%) is even larger. However, the largest gap is amongst non-poor households (41.82%) and poor households (9.14%).

Table 3 provides descriptive statistics of our sample. The average income per capita is VND 91.33 million (or equivalently \$3940) per annum. But the income inequality is quite large, with the highest income being 205 times as much as the average level. Seventy four percent of households are headed by male, and 17.3% of them belong to ethnic minority groups. A typical household head spent roughly 8 years in school and is 53 years old. More than six percent of family heads are uneducated, 149 of them have master's degrees, and 28 doctoral degrees. Two heads are orphans, just 13 and 16 years old. Seventy nine percent of heads are living with their spouses, 44% are members of local associations, and 38% are using internet. Thirty percent of households reside in urban areas, and 17% could borrow money from formal financial institutions. Households have rights to use on average 5,789 m<sup>2</sup> of cultivated land, most of which is in rural areas. They have a mean number of 3.7 members, 2.8 of whom are income-earners, specifically, 1.2 in farming sectors, 1.08 in non-farming sectors, and 0.48 self-employed. Ten percent of households were officially classified 'poor' in 2017

**Table 4.** Bayesian Estimation Results of Eq. (1)

	Mean	Std Dev	MCSE	Median	Equal-tailed	
					[95% Cred.	Interval]
Male	-0.0290	0.0011	0.0003	-0.0290	-0.0310	-0.0269
Ethnic minority	-0.2712	0.0009	0.0002	-0.2712	-0.2731	-0.2693
Number of schooling years	0.0379	0.0006	0.0000	0.0379	0.0366	0.0391
Age	0.0304	0.0003	0.0000	0.0304	0.0298	0.0309
Age <sup>2</sup>	-0.0003	$3.9 \times 10^{-6}$	$3 \times 10^{-7}$	-0.0003	-0.0003	-0.0003
Marital Status	0.0494	0.0013	0.0004	0.0497	0.0470	0.0515
Urban areas	0.1872	0.0008	0.0002	0.1871	0.1859	0.1886
Poverty status in 2017	-0.5578	0.0053	0.0015	-0.5598	-0.5653	-0.5491
Household size	-0.0935	0.0004	0.0000	-0.0935	-0.0943	-0.0927
Cultivated land area (thousand m <sup>2</sup> )	0.0023	0.0001	$5.3 \times 10^{-6}$	0.0023	0.0020	0.0025
Number of farming workers	-0.0581	0.0004	0.0001	-0.0580	-0.0588	-0.0574
Number of non-farming workers	0.2084	0.0013	0.0003	0.2083	0.2061	0.2112
Number of self-employed	0.3371	0.0007	0.0001	0.3372	0.3359	0.3382
Formal credit	0.0153	0.0008	0.0001	0.0152	0.0139	0.0168
Subsidies for poor (VND mil)	0.0061	0.0007	0.0001	0.0061	0.0047	0.0075
Subsidies for non-poor (VND mil)	0.0002	0.0005	0.0001	0.0001	-0.0006	0.0009
Local association membership	-0.0023	0.0008	0.0002	-0.0023	-0.0037	-0.0010
Internet access	0.2938	0.0004	0.0000	0.2938	0.2930	0.2945
Unfair commune	-0.0223	0.0022	0.0007	-0.0228	-0.0260	-0.0186
$\sigma^2$	0.3158	0.0021	0.0000	0.3158	0.3117	0.3201
Acceptance rate	0.3043					
Number of observations	43,093					

Notes: Coefficients for regional dummies and the constant are not reported.

under the national criteria. They receive an average subsidy of VND 5.07 million, 8.6 times as much as that of VND 586,314 for non-poor households. Even though poor families are eligible for such grants, some in 171 communes do not get them.

We estimate Eq. (1) by the Bayesian approach, i.e., the random walk Metropolis-Hastings (MH) Markov chain Monte Carlo (MCMC) method. It employs the normal priors with 0 mean and variance of 10,000 for the regression coefficients. The first 2,500 burn-in iterations are discarded and the subsequent 10,000 MCMC iterations are used to produce the results that are presented in Table 4. The first column shows the posterior mean estimate, the second column the estimated posterior standard deviation, the third column the Monte Carlo standard error (MCSE) measuring the accuracy of simulation results, the fourth column the posterior median estimate, and the last two columns the 95% equal-tailed credible interval.

Other things held constant, income per capita in a male-headed household is 2.9% lower than that in a female-headed one. An ethnic minority family has income per person equal to 76% of that in a Vietnamese or Chinese family. Each additional year spent in school by head would increase his/her household's welfare by 3.9%. This average rate of return to education fits in the range estimated by McGuinness et al. (2021) for Vietnam. Income per capita would increase by

5.1% if the head is living with his/her spouse, and by 20.6% if the household dwells in urban areas. Earnings tend to exhibit an inverted U-shaped motif, rising with age, reaching their peak when heads are 56 years old, then dropping slightly as heads enter retirement. As the number of persons in a household rises by one, their income per capita decreases by 8.9%.

Over the past three decades, Vietnam has experienced one of the most rapid structural transformations among low-income agricultural countries. Massive expansion of the non-farming sectors induced by urbanization and industrialization has created numerous job opportunities with better salaries, and moved millions of young workers out of farming (McCaig and Pavcnik, 2017). Even though larger arable land area is associated with a higher likelihood to use mechanization and to receive credit, its value has become rather low. Table 4 shows that another thousand square meters of cultivated land would raise income per capita marginally by 0.2%, other things being fixed. The relative decline of agriculture in Vietnam is also reflected in the impacts of the numbers of workers on households' well-being. While an extra self-employed or non-farming employee would increase income per capita by 33.7 or 20.8%, an extra farming employee would decrease it by 5.8%.

Being able to borrow money from a formal financial institution gives borrowers a clear advantage because the interest rates are either lower or controlled, loan term length is usually longer than informal credit. Borrowing households can make necessary investments to reallocate their productive resources into more efficient uses. Their income per capita is unsurprisingly 1.5% higher. However, formal credit normally requires collateral that many poor households lack. This and other inherent weaknesses deprive people in households that were labeled 'poor' in the previous year of 42.8% of what they would have if they live in non-poor households. Therefore, government subsidies are a considerable support for them. Additional VND 1 million of grants would augment income per capita in poor households by 0.6%, and not affect non-poor households. The contradicting impacts justify bigger subsidies for the poor and a firm assurance that this policy should be implemented properly. Unfair communes that fail to achieve complete implementation would reduce income per capita of both poor and non-poor households by 2.2%.

Estimation results confirm our preliminary notice in Table 1. Membership of local associations would decrease income per capita slightly by 0.2% whereas internet access would increase it by 34.1%. The emergingly strong influence of internet is seemingly attributed to widespread introduction of high speed connection and increasingly rich information in Vietnamese available on internet. The government should consider ways to upgrade the capabilities of local associations which are expected to play an important part in helping small agricultural producers overcome challenging barriers to participate in global supply chains.

## 5 Conclusion

This paper examines the roles of associations and government at grassroots level and internet access in households' income in Vietnam. We use the random walk

Metropolis-Hastings Markov chain Monte Carlo method with data from the Vietnam Household Living Standards Survey in 2018. It seems that internet usage has made a profound impact on people's well-being. In contrast, it does no good to participate in local associations which even marginally reduce their members' income per capita. Unfair communes that do not give subsidies to poor households harm not only poor but also non-poor families living in these communes. Therefore, the government should facilitate internet access and post more instructive videos in Vietnamese language online. In addition, comprehensive evaluation indicators on the performance of grassroots-level governments should be constructed in order to enforce them to follow national laws and policies, and to improve their efficiency.

## Conflict of interest

The authors declare that they have no conflict of interest.

**Acknowledgments.** Chon Van Le acknowledges financial support from the Vietnam National University, Ho Chi Minh City under the research project B2022-28-05. The authors are very grateful to anonymous referees for their valuable comments.

## References

- Campos, N.F., Nugent, J.B.: Development performance and the institutions of governance: evidence from East Asia and Latin America. *World Dev.* **27**(3), 439–452 (1999)
- Cole, S.A., Fernando, A.N.: The value of advice: evidence from mobile phone-based agricultural extension. Harvard Business School Working Paper (2012). No. 13-047
- Fu, X., Akter, S.: The impact of mobile phone technology on agricultural extension services delivery: evidence from India. *J. Dev. Stud.* **52**(11), 1561–1576 (2016)
- Gupta, S., Davoodi, D., Alonso-Terme, R.: Does corruption affect income equality and poverty? *Economics Governance* **3**(1), 23–45 (2002)
- Jensen, R.: The digital divide: information (technology) market performance, and welfare in the South Indian fisheries sector. *Q. J. Econ.* **122**(3), 879–924 (2007)
- Kaufmann, D., Kraay, A., Zoido-Lobaton, P.: Governance Matters II: Updated Indicators for 2000/01. World Bank Policy Research Department Working Paper No. 2772. World Bank, Washington, DC (2002)
- Kaufmann, D., Kraay, A., Mastruzzi, M.: Governance Matters VIII: Aggregate and Individual Governance Indicators. World Bank Policy Research Department Working Paper No. 4978. World Bank, Washington, DC (2009)
- Khan, M.H.: Governance, Economic Growth and Development since the 1960s. DESA Working Paper No. 54ST/ESA/2007/DWP/54. Department of Economic and Social Affairs, United Nations, New York, NY (2007)
- Krueger, A.O.: The political economy of the rent-seeking society. *Am. Econ. Rev.* **64**(3), 291–303 (1974)
- Lassen, D.D.: The effect of information on voter turnout: evidence from a natural experiment. *Am. J. Polit. Sci.* **49**(1), 103–118 (2005)

- Leng, C., Ma, W., Tang, J., Zhu, Z.: ICT adoption and income diversification among rural households in China. *Appl. Econ.* **52**, 3614–3628 (2020). <https://doi.org/10.1080/00036846.2020.1715338>
- Lenka, S.K., Barik, R.: Has expansion of mobile phone and internet use spurred financial inclusion in the SAARC countries? *Financ. Innov.* **4**, 1–19 (2018). <https://doi.org/10.1186/s40854-018-0089-x>
- Markelova, H., Meinzen-Dick, R., Hellin, J., Dohrn, S.: Collective action for smallholder market access. *Food Policy* **34**(1), 1–7 (2009)
- Mauro, P.: Corruption and growth. *Q. J. Econ.* **110**(3), 681–712 (1995)
- McCaig, B., Pavenik, N.: Moving out of agriculture: structural change in Vietnam. In: McMillan, M., Rodrik, D., Sepulveda, C.P. (Eds.) *Structural Change, Fundamentals, and Growth: A Framework and Case Studies*, International Food Policy Research Institute, Washington, D.C. (2017)
- McGuinness, S., Kelly, E., Pham, T.T.P., Ha, T.T.T., Whelan, A.: Returns to education in Vietnam: a changing landscape. *World Dev.* **138**, 105205 (2021). <https://doi.org/10.1016/j.worlddev.2020.105205>
- Munda, G.: *Dealing with fairness in public policy analysis*. EUR 28751 EN. Publications Office of the European Union, Luxembourg (2017). <https://doi.org/10.2760/75185>. ISBN: 978-92-79-72292-9
- Narayan, D., Pritchett, L.: Cents and sociability: household income and social capital in rural Tanzania. *Econ. Dev. Cult. Change* **47**(4), 871–897 (1999)
- Nguyen, C.V., Giang, L.T., Tran, A.N., Do, H.T.: Do good governance and public administration improve economic growth and poverty reduction? the case of Vietnam. *Int. Public Manage. J.* **24**(1), 131–161 (2021)
- North, D.C.: The new institutional economics and third world development. In: Harris, J., Hunter, J., Lewis, C.M. (Ed.) *The New Institutional Economics and Third World Development*, vol. 21. Routledge, London (1995)
- Ogutu, S.O., Okello, J.J., Otieno, D.J.: Impact of information and communication technology-based market information services on smallholder farm input use and productivity: the case of Kenya. *World Dev.* **64**, 311–321 (2014)
- Organisation for Economic Cooperation and Development (OECD), *OECD Annual Report*. OECD Publishing, Paris (2006)
- Rodrik, D., Subramanian, A., Trebbi, F.: Institutions rule: the primacy of institutions over geography and integration in economic development. *J. Econ. Growth* **9**(2), 131–165 (2004)
- Stiglitz, J.E.: *Globalization and Its Discontents*. W. W. Norton & Company, New York (2002)
- Tanzi, V., Davoodi, H.R.: *Corruption, Growth, and Public Finances*. IMF Working Paper No. 00/182. International Monetary Fund (IMF), Washington, DC (2000)
- Tadesse, G., Bahiigwa, G.: Mobile phones and farmers marketing decisions in Ethiopia. *World Dev.* **68**, 296–307 (2015)
- Tavares, J., Wacziarg, R.: How democracy affects growth. *Eur. Econ. Rev.* **45**(8), 1341–1378 (2001)
- Tran, T.Q., Doan, T.T., Vu, H.V., Nguyen, H.T.: Heterogeneous impacts of provincial governance on household welfare in Vietnam. *Int. J. Soc. Welfare* **28**(2), 229–240 (2019)
- United Nations Development Programme (UNDP), *Social Services for Human Development: Vietnam Human Development Report: 2011*. UNDP, Hanoi, Vietnam (2011)
- World Bank (WB), *The East Asian Miracle: Economic Growth and Public Policy*. Oxford University Press, New York (1993)





# Predicting the Impact of Covid Pandemic on the Relationship Between Logistics Activities and Business Performance: A PLS-SEM Approach

Le Thi Phuong Thanh<sup>(✉)</sup>, Le Thi Phuong Thao<sup>(✉)</sup>, and Tong Viet Bao Hoang<sup>(✉)</sup>

Faculty of Business Administration, University of Economics, Hue University, Huế, Vietnam  
{lthiphuongthanh, ltpthao, tvbhoang}@hueuni.edu.vn

**Abstract.** The purpose of this paper is to examine the relationship between logistics activities and business performance in the Covid-19' impact. A literature review on supply chain and logistics was conducted to integrate the existing knowledge of supply chain management, logistics activities and critical element for the success of business process. A set of logistics activities, business performance based on balance score card, and a conceptual framework are presented. The model was validated through a survey of 212 survey samples at textile enterprises. Partial least squares structural equation modeling (PLS-SEM) was conducted to assess the status of the impact of logistics activities on business performance of textile enterprises in Binh Tri Thien area in the context of the global economic crisis current Covid-19 pandemic.

According to the findings, the logistics activities had impact on the business performance. Especially, the moderator variable "Covid" showed a negative impact on logistics activities and business performance. Finally, based on the current situation, the study has proposed policy implications to improve Logistics activities, thereby contributing to the future business performance of textile enterprises in Binh Tri Thien.

**Keyword:** Logistics activities · Business performance · PLS-SEM

## 1 Introduction

In the current context, the COVID-19 pandemic has strongly impacted the economies and social life of the whole world, upsetting the global supply chain, including logistics activities. When the global market develops with technological advances, especially the opening of markets in developing countries such as Vietnam, logistics is considered by managers as a tool and means for implementation strategic goals of the enterprises. Today's logistics activities are not only associated with warehouse activities, freight forwarding, but also planning and arranging the flow of raw materials and materials from suppliers to manufacturers, then moving goods from production to the final consumer,

creating a connection throughout society in ways that optimize, reduce transportation and storage costs. A well-executed logistics chain will help solve both inputs and outputs for businesses effectively, three products and services to the right place, at the right time with minimal costs while still satisfy the requirements of society and consumers. Therefore, logistics activities have, are and will have a great influence on the business performance of enterprises.

In Vietnam, textile is the second largest industry in terms of import and export turnover. Therefore, the logistics industry plays an important role in the import and export of goods, raw materials, and materials of enterprises in the process of global economic integration. It can be said that the cost of transportation, warehousing and storage of goods, raw materials, etc.... has created great pressure on large-scale textile manufacturing enterprises participating in the export market and participate in the global supply chain (2021). However, at present, Vietnam's logistics industry still has many limitations and difficulties. According to the World Bank's ranking, in 2016, logistics in Vietnam ranked 64/160 worldwide, down compared to 2015. In the ASEAN region, Vietnam ranked 4th after Singapore, Malaysia, and Thailand. Most of Vietnam's logistics enterprises are domestic with small scale, single service, still very fragmented, there is no linkage to create a service chain between businesses, mainly just basic freight forwarding, bear high costs, depend on service designations of foreign economic organizations with the service rate up to 30%; scope of operation is still small. In addition, human resources are not standardized, technology application is at a low level, and there is a serious shortage of human resources with logistics management qualifications as well as foreign language skills. In this general context, Binh Tri Thien area in the map of textile production of the North Central region has a large concentration of enterprises. In the future, there will be a key investment for textile enterprises in this area. Therefore, solving the problem of logistics activities for this area plays an important role in creating effective business activities or not. Especially in the context of the widespread Covid-19 pandemic and global outbreak, this issue is of more interest to organizations and businesses to cope with difficulties caused by the epidemic. At the same time, increasing operational efficiency as well as saving costs.

## 2 Literature Review

Logistics is described as activities (services) related to logistics and transportation, including jobs related to supply, transportation, production tracking, warehousing, distribution procedures, customs, so logistics is a collection of activities of many industries and stages in a complete process (Liyan Zuo, 2016), (Doan Thi Hong Van, 2010). Thus, it can be understood that logistics are services related to activities that ensure the optimization of the entire production and business process, including from input supply to product consumption, which are self-organized by enterprises, implemented or outsourced which has an impact on the business performance of each enterprise in the market (Đặng Đình Đào, 2003) (Lê Công Hoa, 2013) (Hu Mingming, 2010). Quality of logistics activities for textile enterprises is constituted by transportation services, supply of raw materials and necessary elements for production, transportation services, distribution of raw materials and finished products from home. Machines to users and other logistics

services such as packaging, delivery, warehousing, procedures, paperwork, information services. Based on classifications of logistics activities in textile enterprises, research the research and identification of logistics activities includes the following activities:

**Internal logistics:** In the flow of segmented materials within the factory, raw materials are transported from the warehouse to the production site, and they are handled during production. This activity is also known as internal logistics. The task of internal logistics is to establish relationships between production departments in the enterprise. Therefore, internal logistics must transport raw materials and semi-finished products to parts in the production process. Effective internal logistics activities will help businesses save costs and increase business efficiency (Nguyễn Đình Hiền, Đặng Đình Đào, 2013) (Lê Công Hoa, 2013) (Trần Văn Hòa, 2014).

**Inbound logistics:** Inbound logistics activities are responsible for transporting and supplying materials and elements needed for production from the supplier to the factory or production site. Inbound logistics is also responsible for regulating the relationship between the company and its material suppliers (Angelisa Elisabeth Gillyard, 2003). The input logistics activities directly affect the revenue, cost, and competitive advantage of the business (Lê Công Hoa, 2013).

**Outbound logistics** is the process of transporting and distributing raw materials and finished products from the factory to the user. Outbound logistics is also responsible for establishing relationships between the company and its customers. In addition to directly affecting costs, revenue, and output logistics, it also directly affects the competitive position of enterprises (Lê Công Hoa, 2013) (Lai, 2002) (K. Kavčič, 2016).

**Support logistics** is the process of transporting and distributing raw materials and finished products from the factory to the user. Support logistics is also responsible for establishing relationships between the company and its customers. In addition to directly affecting costs, revenue, and output logistics, it also directly affects the competitive position of enterprises.

*Logistics costs:* Whether outsourcing or self-service, businesses always must incur costs for logistics activities, which may include transportation costs, inventory costs, warehousing costs, order processing costs. Goods and information systems, ordering costs, etc. (Yuehua Yuan, 2018). The cost level will directly affect the business performance of textile enterprises. The more reasonable the cost of logistics activities, the higher the business efficiency and vice versa (Nguyễn Xuân Hào, 2015).

### **The Relationship of Logistics Activities with Business Performance of Enterprises**

Manufacturing enterprises in general and textile enterprises create products supplied in the market to make profits. Therefore, logistics activities also have a close relationship in improving the competitiveness and operational efficiency of these enterprises. Especially in improving the competitiveness and operational efficiency of enterprises in specific markets:

Logistics activities improve production, use rationally, save resources, reduce costs for the production process, improving the competitiveness of enterprises (Angelisa Elisabeth Gillyard, 2003) (Krauth, 2005).

Logistics activities ensure supply at the right time and place, helping the production process to go smoothly flow, contributing to improving product quality and lowering product prices (Krauth, 2005) (Lai, 2002).

Logistics activities help managers make decisions about the source of raw materials, the quantity to be supplied and the optimal time to minimize the costs incurred, ensuring the efficiency of production and business activities (production and business) (Angelisa Elisabeth Gillyard, 2003) (Krauth, 2005) (Lai, 2002).

Logistics activities contribute to increase the business value of enterprises through the implementation of additional circulation services (services that continue the production process in the distribution and circulation stage). Logistics is a service with a much larger scale and complexity than pure transportation and forwarding (Angelisa Elisabeth Gillyard, 2003). In the past, the freight forwarder only provided customers with simple, pure, and individual services. Today, due to the development of production and circulation, the details of a product can be supplied by many countries, and conversely, a product of an enterprise can be consumed in many countries, many markets. Different markets, so the services that customers require from distributors, transport and forwarding businesses must be diverse (Krauth, 2005).

### 3 Methodology

The study is based on two groups of secondary data and primary data to analyze the impact of logistics activities on business performance of textile enterprises in Binh Tri Thien area. For secondary data collected from two main sources, which are data sources inside the enterprise and data outside the enterprise such as relevant scientific works, books, internet, etc. collected by drawing on questionnaires and interviews with textile enterprises in Binh Tri Thien area. This study conducted with 212 samples include officials involved in logistics activities at textile enterprises in Binh Tri Thien such as directors, deputy directors, chief – deputy – team leader – staff in charge of logistics activities in textile enterprises.

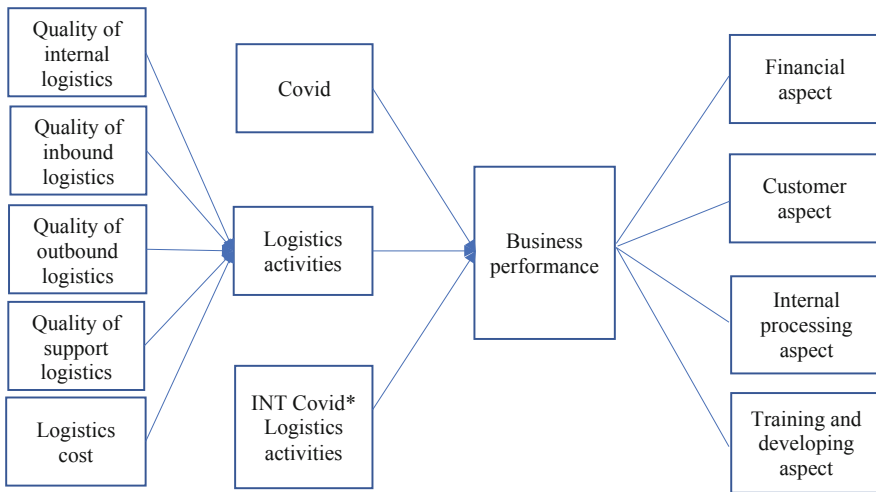


Fig. 1. Proposed research model

For data analysis, the authors used Smart-PLS 3.0 software. The PLS-SEM model is used for data analysis. This method allows for the simultaneous estimation of multiple equations as well as factor analysis and regression analysis in a single step (Hair et al. 2019). Indicators of outer loading greater than 0.5 and average variance extracted (AVE) greater than 50% will be used to test for the convergence of factors.

According to Hair et al. (2019), the factor meets the requirement of reliability as measuring the path through the observed variables when the composite reliabilities (CR) were greater than 0.7 and the Cronbach's Alpha coefficient was greater than 0.7. Furthermore, Fornell and Larcker (1981) stipulate that the factors be completely distinct. The discriminant test is based on AVE having a quadratic value that is greater than the correlation coefficients.

Concerning the scale, the research has summarized in-depth studies on the impact of Logistics activities on production and business activities by authors Nguyen Xuan Hao (2015). Logistics activities have an impact on the efficiency of production and business activities in the textile industry. Combining studies on the use of the Balance Score Card with the scale of (Beata Sadowska, 2015), Nguyen Minh Tam (2009) and Le Thi Phuong Thao (2016), in which the author added the criterion "The electronic document management process is upgraded", "Effective investment project management process" to identify the elements of business performance. The relationship between Logistics activities and business performance also considers under the impact of Covid-19 Pandemic in the model in Fig. 1:

H1: Quality of internal logistics has a positive relationship with logistics activities of textile enterprises.

H2: Inbound logistics quality has a positive relationship with logistics activities textile enterprises.

H3: Quality of outbound logistics has a positive relationship with logistics activities of textile enterprises.

H4: Quality of support logistics has a positive relationship with logistics activities of textile enterprises.

H5: Logistics costs have a positive relationship with logistics activities of textile enterprises.

H6: Logistics activities have a positive relationship with business performance of textile enterprises.

H7: Covid-19 Pandemic impact on the relationship between logistics activities and business performance of textile enterprises.

H7.1: Covid-19 Pandemic impact negative on business performance of textile enterprises.

H7.2: Covid-19 Pandemic \* Logistics activities impact negative on business performance of textile enterprises (Tables 1 and 2).

**Table 1.** Summary of factors of logistics activities affecting the business performance

Encode	Group of factors	Source
Inter-Log	Quality of Internal Logistics	Le Van Bay, Doan Thi Hong Van
In-Log	Quality of Inbound Logistics	Le Van Bay, Doan Thi Hong Van, Nguyen Xuan Hao
Out-Log	Quality of Outbound Logistics	Le Van Bay, Doan Thi Hong Van, Nguyen Xuan Hao, Angelisa Elisabeth Gillyard
Sup-Log	Quality of Support Logistics	Le Van Bay, Doan Thi Hong Van, Nguyen Xuan Hao, Angelisa Elisabeth Gillyard
Cost-Log	Cost logistics	Le Van Bay, Doan Thi Hong Van, Nguyen Xuan Hao, Angelisa Elisabeth Gillyard
Log-act	Logistics activities	

Source: Compiled by authors

**Table 2.** Business performance by using balance score card

Encode	Creteria	Source
<b>BP</b>	<b>Business Performance</b>	Nguyen Minh Tam (2009) and Le Thi Phuong Thao (2016), (Hu Mingming, 2010), (Beata Sadowska, 2015),
BP1	Financial aspect	
BP2	Customer aspect	
BP3	Internal process aspect	
BP4	Training and development aspect	

Source: Compiled by authors

## 4 Result and Discussion

### 4.1 Descriptive Statistics of Respondents

According to the survey results, the surveyed enterprises are divided into three types: private enterprises, limited liability companies, and joint stock companies. There were 212 survey subjects in total, with limited liability companies accounting for 67.5% of the total sample. This is also consistent with the representativeness because limited liability companies account for a large proportion of the provinces' overall SMEs. The survey subjects are distributed across three provinces: Thua Thien Hue, Quang Tri, and Quang Binh. According to the survey results, 60.8% of the enterprises belong to Thua Thien Hue. The remaining two provinces have roughly the same number of respondents (40%).

The gender of the survey respondents is primarily male, accounting for 71.1% of the total sample, when it comes to the characteristics of the director and management level of the logistics department of the textile enterprises in the Binh Tri Thien area participating in the survey. The seniority of the surveyed subjects ranges from 5 to 10 years, accounting for 46.2% of the total. As a result of the analysis of the sample's enterprise characteristics,

we can see that the sample is quite similar to the general characteristics of the population and is highly representative (Table 3).

**Table 3.** Statistics of respondents

Criteria		Amount	Proportion (%)
<b>Type of enterprise</b>	Co., Ltd	143	67.5
	Joint Stock Company	48	22.6
	Private enterprise	19	9.0
	Other	2	0.9
<b>Province</b>	Thua Thien Hue	129	60.8
	Quang Tri	43	20.2
	Quang Binh	40	18.9
<b>Seniority</b>	Less than 5 years	48	22.6
	From 5 to 10 years	98	46.2
	Over 10 years	66	31.2
<b>Gender</b>	Male	157	71.1
	Female	55	28.9
<b>Total</b>		<b>212</b>	<b>100</b>

Source: Processing survey data

## 4.2 Measurement Model Test

In SMARTPLS, the outer loading is defined as the square root of the absolute value of  $R^2$  from the parent latent variable to the child observed variable. According to Hair et al. (2016), the outer loading factor must be greater than or equal to 0.708 observed variables of quality. According to Hair et al., a sub-observed variable is rated as quality if the parent latent variable explains at least 50% of the variation in that observed variable. According to appendix 1, all indexes are greater than 0.70, so it is suitable.

Individual reliability (factorial loads and commonality) and internal consistency (composite reliability (CR), convergent validity via the Average Variance Extracted (AVE), and finally discriminant validity via the Heterotrait-Monotrait ratio) will be examined in the reliability and validity analysis.

The results show that the convergence of the factors showed that all the factors reached the convergence value with outer loading coefficients greater than 0.5 and AVE values greater than 50%. Furthermore, Cronbach's Alpha is greater than 0.7, and CR is greater than 0.7, indicating that all factors are reliable (Hair et al. 2019). Table 4 contains the specifics.

With the HTMT index, Garson (2016) suggests that the discriminant value between the two latent variables is guaranteed when the HTMT index is less than 1. Henseler

**Table 4.** Validity and reliability of measurement model

	Cronbach’s Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Log cost	0.924	0.925	0.941	0.726
Covid	1	1	1	1
Outbound Log	0.916	0.918	0.938	0.751
Inbound Log	0.895	0.896	0.92	0.657
Support Log	0.922	0.922	0.939	0.719
INTCovid_LogAct	1	1	1	1
Internal Log	0.853	0.853	0.911	0.773
Business performance	0.858	0.865	0.903	0.701
Logistics Activities	0.967	0.968	0.97	0.551

Source: Processing survey data

**Table 5.** Heterotrait-Monotrait Ratio (HTMT)

	Cost Log	Covid	Out-Log	In-Log	Sup-Log	INTCovid_ LA	Inter Log	BP	LA
Cost Log									
Covid	0.039								
Out-Log	0.771	0.027							
In-Log	0.824	0.059	0.848						
Sup-Log	0.687	0.048	0.727	0.843					
INTCovid_ LA	0.427	0.195	0.357	0.387	0.409				
Inter Log	0.804	0.027	0.773	0.825	0.693	0.292			
BP	0.494	0.027	0.534	0.56	0.647	0.33	0.515		
LA	0.848	0.046	0.85	0.862	0.809	0.425	0.705	0.613	

Source: Processing survey data

et al. (2015) suggest that if this value is below 0.9; the value is less than 1. Discrimination will be guaranteed. Meanwhile, Clark & Watson (1995) and Kline (2015) use a more stringent threshold of 0.85.

The Heterotraitmonotrait ratio was used to test the discriminant validity (Ringle et al. 2015). All the ratios demonstrated good discriminant validity (Table 5). The measurement model’s output indicates that various validity and reliability criteria were met. As a result, the constructs and measures could be sufficiently discriminated and appropriated to predict relevance for the structural model and associated hypotheses.



### 4.3 Structure Model Assessment

The Variance Inflation Factor is used to assess the phenomenon of multicollinearity in the structural model (VIF). If the variance exaggeration factor does not exceed 5, the model is thought to be non-collinear. Statistical results show that all variables have  $VIF < 5$ , so there is no multicollinearity phenomenon. The assessment of collinearity is the first step in the structural model analysis. The procedure is required to ensure that the path coefficients calculated by regressing endogenous variables on the attached exogenous variables are not skewed. There are collinearity issues between the exogenous and endogenous variables (Lowry and Gaskin 2014).

**Table 6.** Inner VIF values

	Inter Log	In-Log	Out-Log	Sup-Log	Cost Log	Covid	LA	BP	INTCovid_LA
Inter Log							2.538		
In-Log							2.069		
Out-Log							2.878		
Sup-Log							2.542		
Cost Log							2.82		
Covid								1.047	
LA								1.221	
BP									
INTCovid_ LA								1.27	

*Source: Processing survey data*

Collinearity issues with latent variables may exist if the variance inflation factor (VIF) value is greater than 5 or less than 0.2, according to Wong (2013). Table 6 shows that there are no collinearity problems with the latent variables.

### 4.4 Hypothesis Testing

The structural model path coefficients of the model were assessed using the bootstrapping procedure. Bootstrapping, according to Hair et al. (2019), is a resampling technique for estimating the standard error without relying on distributional assumptions. The bootstrap result approximates data normality. It is used to determine the significance of the t statistic for path coefficients (Wong 2013). The path coefficients determined by the bootstrapping process had significant values in Table 7 and model connections are depicted in Table 7.

Based on the change in the beta value of the model and the corresponding standard error between the original model and the Bootstrap test results, the Bayes Factor is

**Table 7.** Final results of the relationship checking of model's constructs

Hypothesis	Relationship	Regression weight	Mean difference	Standard error	Bayes factor	Results
H1	Internal Logistics->Logistics activities	0.132	0.080	0.024	6.19	Supported
H2	Inbound Logistics->Logistics activities	0.249	0.011	0.034	6.33	Supported
H3	Outbound Logistics->Logistics activities	0.227	0.083	0.026	4.23	Supported
H4	Support Logistics->Logistics activities	0.272	0.060	0.018	4.65	Supported
H5	Cost Logistics->Logistics activities	0.26	0.120	0.038	5.52	Supported
H6	Logistics activities->Business performance	0.524	0.078	0.025	3.24	Supported
H7	H7.1 Covid->Business performance	-0.044	0.090	0.027	6.95	Supported
	H7.2 Covid * Logistics activities ->Business performance	-0.069	0.073	0.023	3.53	Supported

Source: *Processing survey data*

calculated using the Bayes Factor calculator of Harry Tattan-Birch et al. (Birch, Brown, West, & Dienes, 2022).

The Bayes Factor indicators are all greater than 3, so it can be concluded that there is an impact between the pairs of variables.

#### **4.5 Discussion About Current Situation of Logistics Activities at Textile Enterprises in Binh Tri Thien Area in the Covid-19 Pandemic**

Internal logistics: Through the statistical results, it shows that most of the internal factors within the enterprise such as the process of controlling the inventory of raw materials and finished products are evaluated from point 3.36 – 4.19 (Table 8). Most enterprises in Binh Tri Thien area lack resources in terms of means of transportation, human resources, and warehouses yard, which serves well for production and business activities. In fact, most textile enterprises use outsourced logistics services. Transporting goods by road still accounts for the largest proportion of textile enterprises in Binh Tri Thien area. According

to experts in the logistics industry in the region, up to 62.7% of businesses answered that they use road transport very often. Next, tri-modal transportation, water transportation. Very few enterprises use rail and air transportation. The level of specialization in the organization of logistics activities at textile enterprises in the three provinces of Quang Binh, Quang Tri and Thua Thien Hue is still high when many enterprises participating in the survey said that they do not have a logistics department or separate supply chain management within the enterprise's organizational structure. In addition, due to the characteristics of most of the textile enterprises in Binh Tri Thien area with small and micro scale, so the level of expertise is not good. As a result, the coordination process between departments in the enterprise will be broken. This shows that human resources specialized in logistics in this area are still seriously lacking and focus on improving training.

**Table 8.** Descriptive statistics on the quality of internal logistics activities

Encode	Indicators	Mean
Inter-Log1	The inventory of raw materials and finished products is reasonable	4.19
Inter-Log 2	The process of coordination between departments is smooth	3.36
Inter-Log 3	Enterprises have adequate transportation, human resources, warehouses, and capital to support production and business activities	3.44

*Source: Processing survey data*

**Inbound logistics:** According to the evaluation of business directors and management levels of logistics activities, the quality of inbound logistics performance criteria is quite low (Table 9). This is the third factor that affects the business performance of textile enterprises. In the past 2 years, managers also explained that due to the impact of the Covid-19 pandemic, that is why the inbound logistics activities of textile enterprises became difficult. Although placing orders was still easy because each company had good contact with the raw material suppliers, the shipping process became difficult and created disruption. This situation occurs because the schedule of ships and port arrivals is cut in the number of trading days. In addition, road transport between provinces is also difficult and prolonged due to the impact of the Covid epidemic. Another fact is that technical errors are still not fixed in the process of dealing with raw material suppliers. The general situation is that enterprises still mainly use traditional transaction methods such as phone/fax and email to exchange information with suppliers & customers. They need to pay more attention to this issue in the near future because inbound logistics is one of the factors to ensure the production process takes place, if this activity is not well organized, it will lead to the failure of the production process break in the supply chain of textile enterprises.

**Outbound logistics:** Through interviews with directors and managers of the enterprise's logistics department, textile enterprises in Binh Tri Thien area mostly carry out processed goods, as well as subsidiaries that perform contracts from the parent company. Therefore, finding partners and finding customers is not a problem that greatly affects the business performance.

**Table 9.** Descriptive statistics on the quality of inbound logistics activities

Encode	Indicators	Mean
In-Log1	The process of transporting raw materials for manufacturing is convenient	2.75
In-Log 2	It is simple to order production materials	3.44
In-Log 3	Time to order and transport materials to ensure production and business progress	3.81
In-Log 4	The information on the catalogs of the raw material suppliers is fully functional	2.66
In-Log 5	The manufacturer has good working relationships with its raw material suppliers	3.47
In-Log 6	In the process of dealing with raw material suppliers, there are few errors in product errors (technical errors, packaging errors, errors in excess or missing quantities, or incorrect product codes...)	2.92

Source: Processing survey data

**Table 10.** Descriptive statistics on the quality of outbound logistics activities

Encode	Indicators	Mean
Out-Log1	The packing and transportation of finished products to partners and sales locations was convenient	3.87
Out-Log2	The finished product's packaging and preservation are excellent	3.67
Out-Log3	Delivery/shipping time to the place of sale is always on time	3.62
Out-Log4	The enterprises have a good relationship with distributors/partners	3.89
Out-Log5	Less errors and confusion in delivery (quantity, product design, location, object, etc.)	3.86

Source: Processing survey data

According to experts, although the Covid situation brings many difficulties in the implementation of the textile supply chain, there are also great opportunities for Vietnamese textile enterprises to receive more orders from partner. Because most of the countries in the textile processing bloc are almost severely affected by the pandemic and cannot afford to maintain. While Vietnam is still quite well controlled and the supply chain is still in operation, the trend of moving textile orders to Vietnam in 2021 and the coming years is very large. This is a great opportunity and brings challenges for textile enterprises. Enterprises focus on further improving the guarantee of on-time delivery and delivery to partners on time. According to the survey, the time to respond to orders and ship to buyers is still at an average level (Table 10). This is a factor that textile enterprises need to consider and have policies to improve.

Support Logistics: This is the factor that has the greatest impact on the business performance of textile enterprises. Through the survey, factors such as: Complete and regularly updated information about customers and partners, Flexibility of the company

in meeting customer needs in terms of product specifications, time delivery, location... High and outsourced logistics services that meet the needs of businesses are evaluated relatively well. However, to improve the efficiency of business operations, it is necessary to pay more attention to further develop the elements of supporting logistics activities, especially to always satisfy customers when they require changes in products, creating Favorable conditions on the delivery location as well as how to shorten the delivery time... Besides these, there are still factors that are undervalued and need special improvement policies. The process of ordering and handling complaints is still not appreciated. According to the 2020 logistics report, the degree of outsourcing logistics activities at enterprises shows that the outsourcing rate is very different between types of logistics services in manufacturing and trading enterprises in Vietnam. Services such as international transportation, domestic transportation, customs declaration, and freight forwarding have the proportion of outsourced enterprises at over 90% of demand, accounting for 43.6% respectively; 33%; 27.7% and 25.5% (Table 11).

**Table 11.** Descriptive statistics on the quality of support logistics activities

Encode	Indicators	Mean
Sup-Log1	Customer and partner information is complete and up to date	3.90
Sup-Log2	Order processing, and complaint handling all need to be done efficiently	3.44
Sup-Log3	Goods delivery and payment procedures are made convenient	3.70
Sup-Log4	The company is very adaptable in terms of product specifications, delivery time, location, and so on	3.91
Sup-Log5	Find professional logistics service providers quickly and easily	3.68
Sup-Log6	Outsourced logistics services meet the needs of enterprises	3.87

*Source: Processing survey data*

On the contrary, there are respectively 44.7%; 41.5% and 32.7% of businesses surveyed said that they outsource logistics management consulting services, purchasing, and warehousing with a low outsourcing rate of less than 10% of demand. In the Binh Tri Thien area, there is also the trend of outsourcing logistics services, but this market still does not develop logistics activities to support businesses, so the search for suppliers' Professional logistics service is also a very difficult problem.

Logistics cost: Warehouse costs are still high due to many reasons. The transport infrastructure system is not synchronous, the connectivity is still limited between sea, railway, and road transport; lack of national and international logistics centers in key economic regions to act as focal points for goods distribution. Secondly, in terms of scale and scope of services, the centers are generally small and mainly serve several enterprises in the industrial zone or a province, city, or city, but have grown to the scale serving a single industry or business. an economic region.

Equipment for storage, loading and unloading is outdated, out of sync, slows down the loading and unloading process, causing damage to goods. In Binh Tri Thien area, the actual situation of investment in infrastructure system is still limited, that's why textile

**Table 12.** Descriptive statistics on the quality of logistics cost.

Encode	Indicators	Mean
Cost-Log1	Shipping cost is reasonable	3.92
Cost-Log2	The cost of raw material inventory is reasonable	3.67
Cost-Log3	Inventory of finished goods is reasonably priced	3.62
Cost-Log4	Order processing and information system costs are reasonable	3.89
Cost-Log5	The cost of freight forwarding, and payment paperwork is reasonable	3.86
Cost-Log6	The cost of renting a warehouse and checking goods is reasonable	3.50

Source: Processing survey data

enterprises are struggling for very high raw material or finished product inventory costs (Table 12).

#### **4.6 Some Policy Implications to Enhance the Impact of Logistics Activities to Improve Business Efficiency of Textile Enterprises in Binh Tri Thien**

Implications for improving the quality of internal logistics: Through the assessment of the current situation, to further improve internal logistics activities in addition to improving resources in terms of capital, warehousing, and human resources, SMEs need to pay attention. Intention to improve inventory management of textile enterprises; Warehouse management skills and operations of managers and team leaders need to be trained professionally and methodically. Besides, it is necessary to re-evaluate the activities of the purchasing department to ensure the appropriate inventory value and to meet the correct production progress. Applying information technology in logistics activities is also an urgent thing and it is necessary to promote this activity at enterprises in Binh Tri Thien area. Electronic data exchange and sharing (EDI) systems need to be widely applied. For textile enterprises, thanks to information technology, it is possible to monitor the status of their import and export goods, and at the same time update legal policies related to logistics activities to properly enforce current laws.

Implications for improving inbound logistics quality: To improve the quality of inbound logistics, SMEs need to do well in planning and managing raw materials, controlling quality in the production process, and building relationships. Close relationship with raw material suppliers. In addition, textile enterprises in Binh Tri Thien area also need to consider the communication process and how to maintain this relationship. Once again emphasizing the use of comprehensive supply chain management technology in the current textile enterprises is very urgent. In the situation that production activities are all affected by the Covid-19 pandemic, it is required by businesses to not miss or lose connection with the shipping information of suppliers. Therefore, investment in technology is necessary in the current context.

Implications for improving outbound logistics quality: To improve the quality of outbound logistics, textile enterprises need to perform well in distribution activities, handle complaints and complaints well, and be flexible in responding to demand. Client.

In addition, in order to improve the output quality, SMEs need to perform well in product distribution activities, build close relationships with distribution intermediaries, promote customer care activities, be flexible in satisfy the needs of customers, etc. Moreover, now businesses must maintain operations in the situation affected by the pandemic. Therefore, the arrangement of workers to meet the requirements to export goods on time, without product defects is also a matter of concern.

Implications for improving the quality of support logistics: Relationships with 3PL service providers need to be strengthened to take advantage of their capabilities in goods declaration, customs relations, and work transfer. It is also responsible for checking import and export documents, related documents, and goods declaration information. To further support this solution, 3PLs are required to improve staff qualifications, knowledge, and capabilities. Require 3PLs to be more responsible in transporting goods, consider the factor of fines for 3PLs if the goods are not delivered on time, as well as damage during transportation.

Implications for reducing logistics costs: Textile enterprises in Binh Tri Thien area only have access to Tien Sa – Da Nang international port, thus increasing transportation costs. An important factor for this solution to be successful is the professional understanding of the management level and the staff of the logistics department. The application and strict enforcement of regulations related to import and export activities avoid risks for goods in the process of circulation, save time and cost of delivery to the factory as well as successful distribution. Products to customers at the right time, at the right place. Import and export goods in the right type of regulations and accurate declaration help businesses avoid mistakes in tax policy, and incur costs for post-customs clearance inspection, if any. Deciding on the method of transporting goods is very important, requiring the management of the logistics department to make the right decisions to ensure a reasonable inventory, but still to promptly meet the raw materials and accessories for the production process, At the same time, still calculating logistics costs between 3PLs reasonably, reducing costs to the lowest level but still receiving good services from 3PL service providers.

## 5 Conclusion

This study has focused on analyzing the impact of logistics activities on business performance of textile enterprises. As a starting point for developing policy implications for fabricated textile enterprise in the Binh Tri Thien area, describe the current state of logistics activities and identify advantages and drawbacks. The research results have practical significance for improving the efficiency of logistics activities of textile enterprises, as well as enterprises in general in the context of the current difficult Covid-19 pandemic. However, because the number of textile enterprises in Binh Tri Thien area is still limited and the enterprises are quite different in size, the study has the conditions to clarify the differences in logistics activities in different regions. This investigation. This will be a suggestion for researchers to implement and implement in future studies.

## Appendix 1

	Cost-Log	Covid	Out-Log	In-Log	Sup-Log	INTCovid_Log-act	Inter-Log	BP	Log-act
Cost-Log2	0.889								
Cost-Log2									0.767
Cost-Log3	0.827								
Cost-Log3									0.726
Cost-Log4	0.852								
Cost-Log4									0.736
Cost-Log5	0.872								
Cost-Log5									0.754
Cost-Log6	0.834								
Cost-Log6									0.758
Out-Log1			0.931						
Out-Log1									0.79
Out-Log2			0.833						
Out-Log2									0.728
Out-Log3			0.856						
Out-Log3									0.754
Out-Log4			0.846						
Out-Log4									0.771
Out-Log5			0.862						
Out-Log5									0.749
covid		1							
In-Log1				0.799					
In-Log1									0.727
In-Log2				0.736					
In-Log2									0.723
In-Log3				0.77					
In-Log3									0.736
In-Log4				0.807					
In-Log4									0.73
In-Log5				0.882					
In-Log5									0.792
In-Log6				0.862					
In-Log6									0.774
BP1								0.811	
BP2								0.813	
BP3								0.851	
BP4								0.872	
Sup-Log1					0.879				
Sup-Log1									0.729
Sup-Log2					0.817				

(continued)



(continued)

	Cost-Log	Covid	Out-Log	In-Log	Sup-Log	INTCovid_Log-act	Inter-Log	BP	Log-act
Sup-Log2									0.707
Sup-Log3					0.848				
Sup-Log3									0.745
Sup-Log4					0.854				
Sup-Log4									0.74
Sup-Log5					0.829				
Sup-Log5									0.738
Sup-Log6					0.858				
Sup-Log6									0.721
Log-act * Covid						1.397			
Inter-Log1							0.876		
Inter-Log1									0.719
Inter-Log2							0.865		
Inter-Log2									0.724
Inter-Log3							0.896		
Inter-Log3									0.721
Cost-Log	0.837								
Cost-Log									0.731

## References

- Angelisa, E.G.: The Relationship Among Supply Chain Characteristics, Logistics and Manufacturing Strategies, and Performance. The Ohio State University (2003)
- Beata Sadowska, A.L.: Logistics costs and balanced scorecard in business management. *Zeszyty Naukowe Uniwersytetu Szczecińskiego* **120**, 92–104 (2015). <https://doi.org/10.18276/epu.2015.120-07>
- BSC equity research: sector report 2021: textile industry outlook (2021)
- Đặng, ĐĐ: Những vấn đề cơ bản về hậu cần vật tư doanh nghiệp. NXB Thống kê, Hà Nội (2003)
- Esper, T.D.: A framework of supply chain orientation. *Int. J. Logist. Manag.* **21**(2), 161–179 (2010)
- Griffis, S.E.: Aligning logistics performance measures to the information needs of the firm. *J. Bus. Logist.* **28**(2), 35–56 (2007)
- Mingming, H., Xiong, W.: Research on performance evaluation of logistics enterprises based on the balanced scorecard. In: International Conference on Intelligent Computation Technology and Automation, pp. 65–68. <https://doi.org/10.1109/ICICTA.2010.241>
- Kavčič, K., Suklan, J.: Outsourcing logistics activities: evidence from Slovenia. *Promet Traffic Transp.* **28**(6), 575–581 (2016)
- Kohn, J.M.: A structural equation model assessment of logistics strategy. *Int. J. Logistics Manag.* **22**(3), 284–305 (2011)
- Krauth, E., Moonen, H.: Performance indicators in logistics service provision and warehouse management – A literature review and framework. RSM Erasmus University (2005)
- Lai, K.-H., Ngai, E.W.T.: Measures for evaluating supply chain performance in transport logistics. *Transp. Res. Part E: Logistics Transp. Rev.* **38**(6), 439–456 (2002)
- Lê, C.H.: Giáo trình quản trị hậu cần (logistics management). NXB Đại học Kinh tế quốc dân, Hà Nội (2013)

- Lê Văn, B.: Giáo trình Quản trị logistics kinh doanh. IESCL (2010)
- Zuo, L., Wang, Y.: Research on the logistics management in the enterprise supply chain system. In: 6th International Conference on Machinery, Materials, Environment, Biotechnology and Computer (MMEBC 2016), pp. 1940–1943 (2016)
- Penteado Marchesini, M.M., Chicarelli Alcantara, R.L.: Logistics activities in supply chain business process: a conceptual framework to guide their implementation. *Int. J. Logistics Manag.* **27**, 6–30 (2016). <https://doi.org/10.1108/IJLM-04-2014-0068>
- Zhao, M., Dröge, C.: The effects of logistics capabilities on firm performance: customer-focused versus information-focused capabilities. *J. Bus. Logistics* **22**(2), 91–107 (2001). <https://doi.org/10.1002/j.2158-1592.2001.tb00005.x>
- Ristovska, N., Kozuharov, S.: The impact of logistics management practices on company's performance. *Int. J. Acad. Res. Account., Finan. Manag. Sci.* **7**(1), 245–252 (2017). <https://doi.org/10.6007/IJARAFMS/v7-i1/2649>
- Hiền, N.Đ., Đào, Đ.Đ.: Xây dựng và phát triển hệ thống logistics ở nước ta theo hướng bền vững. NXB Lao động – Xã hội, Hà Nội (2013)
- Hào, N.X.: Tác động dịch vụ logistics đến hiệu quả hoạt động kinh doanh của các doanh nghiệp sản xuất trên địa bàn tỉnh Quảng Bình. Đại học Kinh tế quốc dân, Hà Nội (2015)
- Wang, P.J., B. a.: On the evaluation of performance of logistic enterprises based on the system of balanced scorecard. *China Bus. Market* **17**, 58–61 (2003)
- Richey, R.G.: Firm technological readiness and complementarity: capabilities impacting logistics service competency and performance. *J. Bus. Logist.* **28**(1), 195–228 (2007)
- Trần, V.H.: Một số vấn đề lý luận và thực tiễn về phát triển dịch vụ logistics ở Việt Nam. NXB Lao động – Xã hội, Hà Nội (2014)
- Yazdanparast, A., Manuj, I.: Co-creating logistics value: a service dominant logic perspective. *Int. J. Logistics Mana.* **21**(3), 375–403 (2010)
- Yuan, Y., Qiao, P.: Research on the influence of logistics outsourcing on logistics. *Adv. Soc. Sci. Educ. Hum. Res.* **199**, 283–287 (2018)



# Consumption Expenditure Comparison Among Vulnerable Households in Thailand

Supanika Leurcharusmee<sup>1</sup>(✉) and Anasree Chaiwan<sup>2</sup>

<sup>1</sup> Center of Human Resource and Public Health Economics, Faculty of Economics,  
Chiang Mai University, Chiang Mai 50200, Thailand

[supanika.1@cmu.ac.th](mailto:supanika.1@cmu.ac.th)

<sup>2</sup> Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand

**Abstract.** This study examines differences in consumption patterns among vulnerable households in Thailand. These include households without home ownership, households with only the elderly, households with disable persons, households with persons with chronic diseases and need assistance in daily living, skipped-generation households, households with a single mother as head of household, and also multiple-vulnerability households. For the estimation, the adaptive lasso method is used to determine whether the type of household vulnerability significantly determines each category of household expenditure. The results show that different types of vulnerable households have different financial burdens. While households with only the elderly have the highest medical and healthcare expenditures, households without home ownership has the highest housing rents, and skipped-generation households have the highest education expenditures. When consider the nonessential expenditures, households with only the elderly, skipped-generation households, and households with a single mother as head of household have lower expenditures on alcohol and tobacco. Moreover, none of the vulnerable households spends more on gambling than the non-vulnerable group. The different spending patterns of vulnerable households could provide guidance to the government in designing policies that meet the different needs of all vulnerable groups.

**Keywords:** Consumption · vulnerable households · poverty · adaptive lasso · variable selection

## 1 Introduction

It is evident that Thai workers face uncertainty and are vulnerable to income shocks even before COVID-19. In 2019, 6.24% of the Thai population or 4.3 million people were living in poverty. That is, consumption expenditure per household member was below the poverty line, which averaged 2,763 baht per

month [14]. Poverty is dynamic and identifying factors that cause individuals or households to be vulnerable to poverty is critical for policy design [3, 5]. To examine the poverty vulnerability situation in Thailand, this study not only examines consumption levels, but also compares consumption patterns of households in different vulnerable household groups. This provides information on which consumption categories claim a high share of income and which consumption categories are low among each type of vulnerable household.

Published papers have shown several factors that cause individuals or households to be vulnerable to face consumption difficulty and falling into poverty, especially in the low- or medium-income countries. Household characteristics, such as household size, household composition and regions. Specifically, larger households with children and elderly living in the rural area are more likely to live in poverty [10, 16, 17]. Demographics of household members, such as gender, age, education and occupation, are also significant determinants of poverty [2, 8, 17]. Moreover, health status and disability also contribute to the risk of poverty [9, 13]. For consumption pattern, households that are vulnerable to poverty are more likely to experience food insecurity and health problem, as well as are deprived of other basic needs such as electricity and child education [4, 7, 22].

For household vulnerability situation in Thailand, the Thai People Map and Analytics Platform (TPMAP) reports statistics for six types of vulnerable households<sup>1</sup>. In August 2020, TPMAP reports the total of 10,754,205 individuals living in 4,104,450 vulnerable households. Among the vulnerable households, there were 7,203 households with unstable housing conditions, 210,887 households with children living in poverty, 443,743 households with elderly who did not receive the Old-Age Allowance, 22,010 households with disability persons who did not receive the Disability Allowance and 7,989 households with chronically ill people<sup>2</sup>. Moreover, 1,472,352 households were vulnerable in more than one dimensions [21]. For the vulnerability and poverty, TDRI [18] examined the impact of vulnerability on poverty and income reduction. The study finds that male-headed households are more probable to living in poverty. This result is consistent with Klasen and Waibel [11]'s finding that, while female-headed households are more vulnerable in many countries, female-headed households in Thailand tend to have higher consumption. In terms of income reduction, households with household heads aged 60 or older and in agriculture and construction sectors are particularly vulnerable to income declines [18].

For consumption patterns, the 2019 household Socio-Economic Survey (SES) of the National Statistics Office (NSO) reports that Thai households spent

<sup>1</sup> Thai People Map and Analytics Platform (TPMAP) is Thailand's data analytics tool to identify people living in poverty and other vulnerable situations for policy designs developed by the Office of National Economic and Social Development Board (NESDB), the National Electronics and Computer Technology Center (NECTEC), the National Science and Technology Development Agency (NSTDA) and the Ministry of Science and Technologies (MOST).

<sup>2</sup> The TPMAP definition of unstable housing conditions follows the Basic Minimum Needs Survey by the Ministry of Interiors (MOI) and the definition of children under poverty follows the definition of the Ministry of Education (MOE).

an average of 20,742 baht per month, of which food, beverages, and tobacco accounted for the largest share (33.9%), followed by housing and household appliances expenses (21.0%) and vehicle related expenses (17.%) [15]. There is limited research on vulnerable and consumption patterns in Thailand. Manajit et al. [12] examines factors affecting consumption patterns in using the 2015 SES data. Using the clustering method, the study classifies households by consumption patterns into five groups, which are food-dominated household, housing-dominated household, food and housing-dominated household, miscellaneous goods and services-dominated household and transportation-dominated household. The study also examines factors determining consumption patterns and found that level of education and age of the head of household and region of residence affect the consumption patterns of all groups. Moreover, households with elders are more likely to be food-dominated households, housing-dominated households and food and housing-dominated households and less likely to be miscellaneous goods and services dominated households. Wongmonta [23] uses 2015 SES data to examine consumption patterns for food, beverage and tobacco across households. The results show that larger households with higher proportions of elderly and children living in rural area are likely to have lower scaled food consumption expenditures.

To examine consumption expenditure patterns among vulnerable households, this study used cross-sectional data from the 2019 Household Socio-Economic Survey (SES) by Thailand's National Statistical Office (NSO). The vulnerable households investigated in this study include households that are vulnerable due to high dependency ratio, health and disability, as well as the lack of assets. Specifically, this study examines households without home ownership, households with only the elderly, households with disable persons, households with persons with chronic diseases and need assistance for daily life, skipped-generation households, households with a single mother as head of household and also multiple-vulnerability households. For the consumption pattern, we compare consumption expenditures among vulnerable and non-vulnerable in three groups of consumption, which are essential expenditures, enhancing-quality-of-life expenditures and nonessential expenditures (e.g. gambling, alcohol and tobacco expenditures). It should be noted that the data are from 2019 and the COVID-19 effect has not been taken into account.

To examine whether households that fall under a type of vulnerability have different expenditures on ten consumption categories compared to non-vulnerable households, statistical significance generally can be tested with several hypothesis testing methods such as ANOVA or simple regressions. However, the standard methods rely on the p-value to indicate the statistical significance. As noted in several studies, the p-value has a disadvantage in that it decreases with sample size [6, 19]. As this study uses the 2019 SES data with the large sample size of 45,586 households, this study adopts the Least Absolute Shrinkage and Selection Operator (Lasso) method, which is a model selection method that is not based on the p-value. Lasso is a regression method developed by Tibshirani [20] with L1-regularization to reduce the number of covariates in order

to improve prediction accuracy. As the standard lasso has the oracle properties only under some restrictive conditions, which can lead to inconsistent selection of variables. Zou [24] then developed the adaptive lasso method, which adds adaptive weights to the standard lasso model to allow different regularization for each coefficient. The method has the oracle properties and improves the variable selection consistency. Therefore, Zou [24]’s adaptive lasso method is used for analysis in this study. Since consumption is a standard measure of poverty and household consumption expenditure can reflect people’s well-being, the results of this study would help policy makers to identify the needs of each vulnerable group.

## 2 Methodology

This study adopts Zou [24]’s adaptive lasso method to investigate the difference consumption pattern among households with different types of vulnerability. For each category of consumption expenditure, the adaptive lasso estimator is as follows:

$$\hat{\beta}_* = \arg \min_{\beta} \left( \left\| y - \sum_{j=1}^p x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j|^\gamma} \right), \quad (1)$$

where  $y$  is the natural log of per-capita household consumption expenditure and  $x$  is a vector of all explanatory variables including dummy variables representing seven types of vulnerability including households without home ownership (V1), households with only the elderly (V2), households with disable persons (V3), households with persons with chronic diseases and need assistance for daily life (V4), skipped-generation households (V5), households with a single mother as head of household (V6), multiple-vulnerability households (V Multiple) and controlled variables. It should be noted that V1–V6 represent households with only one type of vulnerability, while V Multiple presents households with at least 2 types of vulnerability. Similar to standard lasso,  $\lambda > 0$  is a tuning variable. The adaptive weight is  $\frac{1}{|\hat{\beta}_j|^\gamma}$ , where  $\hat{\beta}$  is the initial value of regression coefficients usually estimated using the ordinary least squared or ridge regression and  $\gamma > 0$  [1].

## 3 Data

This study uses data from the 2019 Socio-Economic Survey (SES) collected by Thailand’s National Statistical Office (NSO). A total of 45,586 households were included in the SES. To compare the consumption patterns of households with different types of vulnerability, we define six types of vulnerability, including households without home ownership, households with only the elderly, households with disable persons, households with persons with chronic diseases and

need assistance for daily life, skipped-generation households, households with a single mother as head of household. Some households have more than one type of vulnerability and are separately classified in the multiple vulnerability group. Therefore, in this study, there is one group of non-vulnerable households and a total of seven types of vulnerability, which are six groups of households with only one type of vulnerability (V1-V6) and one group of households with multiple vulnerabilities (V Multiple), as shown in Table 1.

For consumption expenditures, we examine three categories and ten sub-categories of per-capita household consumption expenditures. Among the three expenditure categories, *Category 1* is expenditures on essential consumption including food, medical care and housing rent. *Category 2* is expenditures on consumption to improve living standards including telecommunications, utilities, education and fuel. Finally, *Category 3* is the expenditures on nonessential consumption including gambling, alcohol and tobacco. The basic statistics are shown in Table 1.

**Table 1.** Variable description and sample statistics

VARIABLES	Description	N	Mean	SD
<i>Vulnerable type</i>				
V1	Households without home ownership (Do not own their current residential property)	45,586	0.20	0.40
V2	Households with only the elderly aged 60 and above	45,586	0.12	0.33
V3	Households with at least one disable person	45,586	0.04	0.19
V4	Households with at least one person with chronic diseases and need assistance for daily life	45,586	0.06	0.24
V5	Skipped-generation households (No working-aged generation)	45,586	0.01	0.09
V6	Households with a single mother as head of household	45,586	0.02	0.15
V Multiple	Households with at least two types vulnerability	45,586	0.11	0.32
<i>Household consumption expenditure (in-cash)</i>				
oexp_fd_incashpc	Per-capita food expenditure	45,586	2,342.00	1,444.00
oexp_med_incashpc	Per-capita medical expenditure	45,586	85.51	396.30
oexp_houserent_incashpc	Per-capita house rent expenditure	45,586	198.40	624.60
oexp_telecom_incashpc	Per-capita telecom and internet expenditure	45,586	267.80	259.80
oexp_utility_incashpc	Per-capita utility expenditure	45,586	359.70	328.10
oexp_edu_incashpc	Per-capita education expenditure	45,586	70.90	333.90
oexp_fuel_incashpc	Per-capita fuel expenditure	45,586	527.00	686.10
oexp_gambling_incashpc	Per-capita gambling expenditure	45,586	93.06	254.80
oexp_alcohol_incashpc	Per-capita alcohol expenditure	45,586	54.11	255.60
oexp_tobacco_incashpc	Per-capita tobacco expenditure	45,586	30.78	134.60
<i>Household characteristics</i>				
Urban	Administrative Area	45,586	0.57	0.50
HHsize	Household size	45,586	2.74	1.51
HHincomepc	Per-capita household income (in-cash)	45,586	8.59	1.30

Note: (1) Calculated from the SES 2019 sample (with no population weighting). (2) The dummy variables V1-V6 only equal to 1 if the household only face one type of vulnerability. (3) The unit of per-capita expenditure and income is baht/month/person.

## 4 Results

This study examines differences in consumption patterns among vulnerable households in Thailand including households without home ownership (V1), households with only the elderly (V2), households with disable persons (V3), households with persons with chronic diseases and need assistance for daily life (V4), skipped-generation households (V5), households with a single mother as head of household (V6) and multiple-vulnerability households (V Multiple).

**Table 2.** Per-capita household expenditures and incomes of households with different types of vulnerability.

	Expenditure (Baht/month)	Income (Baht/month)	Ratio of expenditure to income (%)
V1	10,553	12,872	82.0%
V2	5,244	7,128	73.6%
V3	4,130	5,176	79.8%
V4	4,717	6,151	76.7%
V5	3,283	3,704	88.6%
V6	4,142	5,173	80.1%
V Multiple	4,872	5,834	83.5%
Non-vulnerable	7,249	9,656	75.1%
<b>All</b>	<b>7,183</b>	<b>9,210</b>	<b>78.0%</b>

Note: (1) The total expenditure includes all in-cash expenditures of both consumption and non-consumption expenditures (such as tax, interest expense, fine, donation, etc.). (2) The income includes all in-cash income from both labor market or business incomes of all household members and other sources of income (such as pension, assistance money, interest income, dividend, etc.). (3) The statistics are calculated using the 2019 SES data with population weighting. It should be noted that some vulnerable groups, namely skipped-generation households and households with a single mother as head of household, have smaller sample size and the statistics may not represent the population.

Overall, the per-capita household expenditures and incomes of households with different types of vulnerability are shown in Table 2. It should be noted that the data are from 2019 and the COVID-19 effect has not been taken into account. From these results, skipped-generation households (V5) had the lowest levels of both income and consumption. Moreover, skipped-generation households also had the higher expenditure-to-income ratio at 88.6% indicating that the skipped-generation households were the most financially vulnerable on the average. In addition, households with disable persons (V3) and households with a single mother as head of household (V6) also had low income and consumption. It should be noticed that households without home ownership (V1) are not necessary vulnerable on the average as their income are higher than that of the non-vulnerable group. However, their expenditure-to-income ratio is higher than the non-vulnerable and several other vulnerable groups.



**Table 3.** Factors determining the expenditures for essential consumption

	Standard lasso		
	ln(food exp)	ln (medical exp)	ln (house rent)
V1	0.1101	-0.0878	4.1391
V2	-0.2670	0.4401	-0.1156
V3	-0.2542	0.0647	0.0767
V4	-0.0379	0.1966	0.1904
V5	-0.2091	0.0966	0.0886
V6	-0.0807	-0.0290	0.1275
V Multiple	-0.2238	0.3119	1.5080
Urban	0.1630	0.0600	0.3871
HHsize	-0.0752	0.1043	-0.0838
HHincome	0.2106	0.1477	0.0674
Constant	5.8927	0.3902	-0.5302
N	45,586	45,586	45,586
MSE (Training)	0.6022	4.7388	3.0557
MSE (Validation)	0.6162	4.8018	3.1858
R-squared (Training)	0.1940	0.0145	0.4943
R-squared (Validation)	0.1913	0.0144	0.4752
	Standard lasso		
	ln(food exp)	ln (medical exp)	ln (house rent)
V1	0.1101	-0.0882	4.1297
V2	-0.2670	0.4394	-0.1249
V3	-0.2542	0.0645	
V4	-0.0379	0.1969	0.1742
V5	-0.2091		
V6	-0.0807		
V Multiple	-0.2238	0.3117	1.4954
Urban	0.1630	0.0599	0.3867
HHsize	-0.0752	0.1040	-0.0816
HHincome	0.2106	0.1475	0.0662
Constant	5.8927	0.3928	-0.5145
N	45,586	45,586	45,586
MSE (Training)	0.6022	4.7389	3.0563
MSE (Validation)	0.6162	4.8017	3.1862
R-squared (Training)	0.1940	0.0144	0.4942
R-squared (Validation)	0.1913	0.0145	0.4751

Note: (1) The expenditures are in-cash per-capita household expenditure for each category of essential consumption. (2) All lasso and adaptive lasso regressions are estimated using 10-fold cross validation to minimize the out-of-sample MSE and the goodness-of-fits are reported for both training and validation subsamples.

To further investigate, this study also adopts the adaptive lasso to estimate the factors affecting three categories and ten subcategories of per-capita household consumption expenditures. The results for Category 1, 2 and 3 expenditures are shown in Table 3, 4 and 5, respectively.

Consider the results for the *Category 1* expenditures on essential consumption. For per-capita household expenditures on food, all types of vulnerable households have different levels of food expenditures compared to the base group, which is the non-vulnerable group. While other types of vulnerable groups including the multiple-vulnerability group have lower food expenditures, households without home ownership (V1) have 11% higher food expenditures than the non-vulnerable group. The group with the lowest food expenditure are households with only the elderly (V2), whose food expenditures are 27% lower than those of the non-vulnerable group.

For medical expenditures, households without home ownership (V1) also show a different pattern compared to other vulnerable groups. Specifically, most vulnerable groups have higher medical expenditures than the non-vulnerable group. However, households without home ownership (V1) have lower medical expenditures and households with a single mother as head of household (V6) do not have significantly different expenditures according to the variable selection of the adaptive lasso model. For the size of the expenditure, households with only the elderly (V2) spend the highest amount on medical treatments and spend 44% higher compared to the non-vulnerable group.

The housing rent category is unique in this analysis because households who do not own their living property are considered to be vulnerable and are classified in the vulnerable groups V1 or V multiple. Therefore, this is reflected in the regression results that households without home ownership (V1) and multiple-vulnerability households (V Multiple) are the groups with the highest expenditures on housing rent. It should be noted that some households who own their properties still reported paying some rent, but on average much less than those who do not own properties.

*Category 2* expenditures report consumption that can improve the standard of living. It can be seen that most vulnerable groups, namely households with only the elderly (V2), households with disable persons (V3), households with persons with chronic diseases and need assistance for daily life (V4) and multiple-vulnerability households (V Multiple), have lower expenditures than the non-vulnerable group. For skipped-generation households (V5) and households with a single mother as head of household (V6), they have children and need to spend more on education. However, when other categories of standard-of-living consumption are considered, they spend less than the non-vulnerable group. In addition, the gaps of the expenditures between vulnerable and non-vulnerable households are higher for the standard-of-living consumption than for food, which is essential consumption.

Similar *Category 1* expenditures, households without home ownership (V1) have a different consumption pattern than other vulnerable groups. Households without home ownership have higher expenditures on telecommunication and education compared to non-vulnerable group. However, they have lower expenditures on utility and fuel.

For *Category 3* expenditures, it can be seen that none of the vulnerable groups has higher expenditures on gambling, alcohol and tobacco than the non-

**Table 4.** Factors determining the expenditures for consumption to improve standard of living

	Standard lasso			
	ln (telecom exp)	ln (utility exp)	ln (education exp)	ln (fuel exp)
V1	0.2140	-0.2648	0.2491	-0.8732
V2	-1.1664	-0.1405	-0.5738	-1.7184
V3	-0.6658	-0.3508	-0.5476	-0.9159
V4	-0.3524	-0.1734	-0.4925	-0.5645
V5	-0.7412	-0.1537	1.4002	-0.9870
V6	-0.2696	-0.1448	1.1306	-0.7490
V Multiple	-0.8933	-0.4129	-0.2677	-1.8348
Urban	0.2892	0.3096	0.1340	0.0936
HHsize	0.0890	-0.0289	0.6475	0.2578
HHincome	0.3983	0.2842	0.0321	0.4536
Constant	1.3506	3.0378	-0.7425	1.1176
N	45,586	45,586	45,586	45,586
MSE (Training)	2.0317	1.3988	3.3768	4.0985
MSE (Validation)	2.0269	1.4517	3.3652	4.1427
R-squared (Training)	0.2417	0.1314	0.2437	0.2104
R-squared (Validation)	0.2410	0.1120	0.2389	0.2113
	Adaptive lasso			
	ln (telecom exp)	ln (utility exp)	ln (education exp)	ln (fuel exp)
V1	0.2140	-0.2648	0.2491	-0.8732
V2	-1.1664	-0.1405	-0.5738	-1.7184
V3	-0.6658	-0.3508	-0.5476	-0.9159
V4	-0.3524	-0.1734	-0.4925	-0.5645
V5	-0.7412	-0.1537	1.4002	-0.9870
V6	-0.2696	-0.1448	1.1306	-0.7490
V Multiple	-0.8933	-0.4129	-0.2677	-1.8348
Urban	0.2892	0.3096	0.1340	0.0936
HHsize	0.0890	-0.0289	0.6475	0.2578
HHincome	0.3983	0.2842	0.0321	0.4536
Constant	1.3506	3.0378	-0.7425	1.1176
N	45,586	45,586	45,586	45,586
MSE (Training)	2.0317	1.3988	3.3768	4.0985
MSE (Validation)	2.0269	1.4517	3.3652	4.1427
R-squared (Training)	0.2417	0.1314	0.2437	0.2104
R-squared (Validation)	0.2410	0.1120	0.2389	0.2113

Note: (1) The expenditures are in-cash per-capita household expenditure for each category of essential consumption. (2) All lasso and adaptive lasso regressions are estimated using 10-fold cross validation to minimize the out-of-sample MSE and the goodness-of-fits are reported for both training and validation subsamples.

vulnerable group, except households without home ownership (V1). Households with only the elderly (V2) have the lowest gambling expenditure, while skipped-generation households (V5) have the lowest alcohol expenditures and households with a single mother as head of household (V6) have the lowest tobacco expenditures.

**Table 5.** Factors determining the expenditures for nonessential consumption

	Standard lasso		
	ln (gambling exp)	ln (alcohol exp)	ln (tobacco exp)
V1	-0.2338	0.3128	0.2765
V2	-0.5162	-0.4263	-0.2083
V3	-0.3474	-0.1187	0.0128
V4	-0.1877	-0.0724	-0.0219
V5	-0.3507	-0.4206	-0.2074
V6		-0.3320	-0.2730
V Multiple	-0.4754	-0.2280	-0.1207
Urban	0.0813	-0.0675	-0.2102
HHsize	0.0452	-0.0041	0.0682
HHincome	0.2200	0.0741	-0.0362
Constant	0.3569	0.0704	1.0005
N	45,586	45,586	45,586
MSE (Training)	5.8888	3.1735	2.9789
MSE (Validation)	5.9884	3.2461	2.9250
R-squared (Training)	0.0241	0.0212	0.0145
R-squared (Validation)	0.0198	0.0197	0.0120
	Adaptive lasso		
	ln (gambling exp)	ln (alcohol exp)	ln (tobacco exp)
V1	-0.2338	0.3159	0.2770
V2	-0.5162	-0.4204	-0.2085
V3	-0.3474	-0.1203	
V4	-0.1877	-0.0797	
V5	-0.3507	-0.4216	-0.2057
V6		-0.3364	-0.2704
V Multiple	-0.4754	-0.2285	-0.1191
Urban	0.0813	-0.0674	-0.2101
HHsize	0.0452		0.0671
HHincome	0.2200	0.0745	-0.0363
Constant	0.3569	0.0544	1.0025
N	45,586	45,586	45,586
MSE (Training)	5.8888	3.1735	2.9790
MSE (Validation)	5.9884	3.2461	2.9249
R-squared (Training)	0.0241	0.0212	0.0145
R-squared (Validation)	0.0198	0.0197	0.0120

Note: (1) The expenditures are in-cash per-capita household expenditure for each category of essential consumption. (2) All lasso and adaptive lasso regressions are estimated using 10-fold cross validation to minimize the out-of-sample MSE and the goodness-of-fits are reported for both training and validation subsamples.

Households without home ownership (V1) has higher expenditures for alcohol and tobacco than the non-vulnerable group. Households with a single mother as head of household (V6) has gambling expenditure that is not significantly different from the non-vulnerable households. In addition, households with disable persons (V3) and households with persons with chronic diseases and need assistance for daily life (V4) have tobacco expenditure that is not significantly different from the non-vulnerable households.

It should be noted that all regression results are controlled for per-capita household income, household size and location (urban or rural). The per-capita household income is added because households in different vulnerable groups have different income, which directly affect their expenditures. Moreover, we also use household size to control for the economies of scale that may occur in some categories of household expenditures, such as food or utility expenditures. In addition, location (urban or rural) is added to control for potential different consumption needs and price level.

## 5 Conclusion and Discussion

From this study, we found that households with vulnerability earn much lower income compared to non-vulnerable households on the average, except in the case of households without home ownership. However, when the expenditure-to-income ratio is considered, households without home ownership has a higher expenditure-to-income ratio than the non-vulnerable group.

Moreover, according to the adaptive lasso estimation, different types of vulnerable households have different financial burdens. While households with only the elderly have the highest medical and healthcare expenditures, households without home ownership has the highest housing rents and skipped-generation households have the highest expenditures on education. With these expenditures, households with only the elderly are left with the lowest expenditure on food and telecommunications. The multiple-vulnerability group faces relative high expenditures on medical and health cares and housing rents, leaving them with low expenditures on utility and fuel.

When consider the nonessential expenditures, i.e., gambling, alcohol, and tobacco expenditures, the results show that households with only the elderly, skipped-generation households, and households headed by a single mother have lower alcohol and tobacco expenditures. However, households without home ownership have significantly higher expenditures on alcohol and tobacco compared to the non-vulnerable group. Gambling expenditures show different results. No type of vulnerable household spends more on gambling than the non-vulnerable group, although the gambling expenditures of households with a single mother as head of household are not significantly different from those of the non-vulnerable group.

The results show different spending patterns among vulnerable households that can help the government design policies that meet the needs of each vulnerable group. For example, households with only the elderly, households with chron-

ically ill people, and households that are multiply vulnerable have a high proportion of spending on medicine and health care. This may reflect that there are dimensions of medical or health care treatment that are not practically covered by the universal health coverage program. In addition, households with disable persons, households with persons with chronic diseases and skipped-generation households face relatively lower income and expenditures. They also have higher poverty rates than other vulnerable groups. These results may reflect the inadequacy of the disability allowance for households with disabilities and households with chronic patients, the old-age allowance for households with only the elderly, and the child support grant for skipped-generation households. Because the process of identifying different types of vulnerable households is subject to the risk of exclusion error, the different consumption needs of different types of vulnerable groups and the lower nonessential expenditures on gambling, alcohol, and tobacco suggest that more cash transfers should be considered because their versatile use can benefit a broader range of vulnerable households.

**Acknowledgment.** This study was supported by National Research Council of Thailand (NRCT) under the Khonthai 4.0 spearhead program.

## References

1. Ahrens, A., Hansen, C.B., Schaffer, M.E.: lassopack: model selection and prediction with regularized regression in Stata. *Stand. Genomic Sci.* **20**(1), 176–235 (2020)
2. Awan, M.S., Malik, N., Sarwar, H., Waqas, M.: Impact of education on poverty reduction (2011)
3. Azam, M.S., Imai, K.S.: Vulnerability and poverty in Bangladesh. *Chronic Poverty Research Centre Working Paper*, vol. 141 (2009)
4. Beasley, L.O., Jespersen, J.E., Morris, A.S., Farra, A., Hays-Grudo, J.: Parenting challenges and opportunities among families living in poverty. *Soc. Sci.* **11**(3), 119 (2022)
5. Chaudhuri, S., Jalan, J., Suryahadi, A.: Assessing household vulnerability to poverty from cross-sectional data: a methodology and estimates from Indonesia (2002)
6. Cho, S.E., Geem, Z.W., Na, K.S.: Predicting depression in community dwellers using a machine learning algorithm. *Diagnostics* **11**(8), 1429 (2021)
7. Drammeh, W., Hamid, N.A., Rohana, A.J.: Determinants of household food insecurity and its association with child malnutrition in Sub-Saharan Africa: a review of the literature. *Current Res. Nutr. Food Sci. J.* **7**(3), 610–623 (2019)
8. Feliciano, D.: A review on the contribution of crop diversification to sustainable development goal 1 “No poverty” in different world regions. *Sustain. Dev.* **27**(4), 795–808 (2019)
9. Groce, N., Kembhavi, G., Wirz, S., Lang, R., Trani, J.F., Kett, M.: Poverty and disability—a critical review of the literature in low and middle-income countries. *Leonard Cheshire Research Centre Working Paper Series*, vol. 16 (2011)
10. Kakwani, N., Subbarao, K.: Poverty among the elderly in Sub-Saharan Africa and the role of social pensions. *J. Dev. Stud.* **43**(6), 987–1008 (2007)
11. Klasen, S., Waibel, H.: Vulnerability to poverty in South-East Asia: drivers, measurement, responses, and policy issues. *World Dev.* **71**, 1 (2015)

12. Manajit, S., Samutachak, B., Voelker, M.: Socio-economic determinants of consumption patterns in Thailand. *Asia-Pac. Soc. Sci. Rev.* **20**(2), 39–51 (2020)
13. Mitra, S., Posarac, A., Vick, B.C.: Disability and poverty in developing countries: a snapshot from the world health survey. World Bank social protection working paper, vol. 1109 (2011)
14. NESDC. The 2019 poverty and inequality report. National Economic and Social Development Council, Bangkok (2019). <http://social.nesdc.go.th/social/Default.aspx?tabid=126&articleType=ArticleView&articleId=243>
15. NSO: The 2019 Socio-Economic Survey Report. NSO, Bangkok (2019)
16. Olinto, P., Beegle, K., Sobrado, C., Uematsu, H.: The state of the poor: where are the poor, where is extreme poverty harder to end, and what is the current profile of the world's poor. *Econ. Premise* **125**(2), 1–8 (2013)
17. Shaukat, B., Javed, S.A., Imran, W.: Wealth index as substitute to income and consumption: assessment of household poverty determinants using demographic and health survey data. *J. Poverty* **24**(1), 24–44 (2020)
18. TDRI. Risk and Social Vulnerability Assessment: Practical Guidelines to Measure Poverty and Social Vulnerability in Thailand. Bangkok: TDRI (2006). <https://tdri.or.th/wp-content/uploads/2012/12/h103.pdf>
19. These, M.S., Ronna, B., Ott, U.: P value interpretations and considerations. *J. Thorac. Dis.* **8**(9), E928 (2016)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996)
21. TPMAP. Vulnerable household statistics. Thai People Map and Analytics Platform (2021). <https://www.tpmmap.in.th/fragile>. Accessed 15 Nov 2021
22. Trotta, G., Gram-Hanssen, K., Lykke Jørgensen, P.: Heterogeneity of electricity consumption patterns in vulnerable households. *Energies* **13**(18), 4713 (2020)
23. Wongmonta, S.: An assessment of household food consumption patterns in Thailand. *J. Asia Pac. Econ.* **27**, 1–21 (2020)
24. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)



# The Presence of Child and Spouse in the Household and Labor Market Opportunities of Male and Female Workers in Thailand

Supanika Leurcharusmee<sup>1</sup>(✉) and Anaspree Chaiwan<sup>2</sup>

<sup>1</sup> Center of Human Resource and Public Health Economics, Faculty of Economics,  
Chiang Mai University, Chiang Mai 50200, Thailand

supanika.l@cmu.ac.th

<sup>2</sup> Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand

**Abstract.** Male and female populations are exposed to different life cycle risks that could reduce labor market opportunities. Since one of the main factors affecting the livelihood of the female population is family and childbearing, this study analyzes the effects of having spouse and child presented in the household on the labor market participation and opportunities of female young adults comparing to male young adults in Thailand. As the estimation faces the crucial problem of selection bias, in which women with higher qualifications or a focus on career success are more likely to choose to remain single or not have children, the Multinomial Treatment Model developed by Deb (2009) is adopted to estimate the effects to correct for the sample selection. The results suggest that the presence of a spouse in the household not only does not reduce the likelihood of female young adults entering the labor market or reduce their incomes, but also leads them to work more and earn higher incomes. However, the presence of children reduces the likelihood that women will enter the labor market, reduce their hours of work, and reduce their incomes. For male young adults, the effects of the presence of a spouse and a child in the household are similar to those for women, except for the labor force participation dimension in the case that both the spouse and the child live together in the household. While female workers living with a spouse and a child are less likely to participate in the labor market, male workers are more likely to work.

**Keywords:** Female labor force participation · Gender income gap · Parenthood penalty · Multinomial Treatment Model · Sample selection

## 1 Introduction

Economic participation and opportunity directly affect the well-being of the population. The male and female populations experience different life cycle risks that could reduce economic opportunities. Some of the risk factors differ between the



male and female populations. As one of the main factors that affect the livelihoods of the female population is family and childbearing factors, the purpose of this study is to analyze the effects of having spouse and child presented in the household on the labor market participation and opportunities of Thai women compared to Thai men.

For an overview of the male-female inequality in Thailand, the World Economic Forum's Global Gender Gap Index measures gender equality in four areas: Health and Survival, Education Attainment, Economic Participation and Opportunity and Political Empowerment. Thailand's gender inequality is moderate. In 2020, Thailand's Gender Gap Index was 0.71 points, ranked 79th in the world. For the Economic Participation and Opportunity subindex, it is measured by the status in the women's labor market including labor force participation, wage equality for similar work, and estimated earned income, ratio of legislators, senior officials and managers and ratio of professional and technical workers.

In 2019<sup>1</sup>, the labor force participation rate for women aged 18–60 was 78.03%. This was lower than the labor force participation rate for men of 90.63%. In terms of income in the labor market, female workers earned an average of 11,864 baht per month, which is lower than the average 13,870 baht per month for male workers. In term of high-skilled positions, the proportion of female workers employed in a professional or technical position was 12.14% of all employed females. This was higher than male workers, where professional or technical positions were accounted for only 7.98% of all employed males. The higher proportion of high-skilled positions of female workers is due to the fact that women in Thailand are more educated than men. Among women aged 18–60, 7.95% had a professional degree and 18.77% had a bachelor's degree or higher, while only 10.73% of men had a professional degree and 12.96% had a bachelor's degree or higher. The gender gap in the participation and economic opportunity sub-index is largely due to the fact that female workers hold fewer management-level positions compared to men. The proportion of female workers in management positions was 2.40% of all employed women. This is less than the male workforce, which accounted for 4.41% of all employed men<sup>2</sup>.

Differences between men and women in labor market participation and opportunities can be due to family factors, such as marriage and childbearing. Over the past 50 years, marriage have been viewed as a choice rather than a practice [10] and the marriage rate falls [4]. Becker [1] developed a rational choice theory to explain marriage and childbearing decisions by comparing costs and benefits. Subsequently, Browning, Chiappori and Weiss [4] outlines the benefits of marriage from an economic perspective as a collaboration for the purpose of mutual production in term of division of labor and consumption. The benefits of

<sup>1</sup> The statistics reported were calculated from the 2019 Socio-Economic Survey (SES) data. The monthly income and proportions of high-skilled positions are the average for all employed workers.

<sup>2</sup> The occupational classification used in this study follows ISCO-08, where Managers are major group 1, Professionals are major group 2 and Technicians and associate professionals are major group 3.

marriage differ across couples and depends on several factors including attitude, culture and economic factors. The main economic factors include education and income [1, 10].

For the decision to have children, Hashemzadeh et al. [12] reviewed 53 studies on factors affecting fertility from 1946 to 2021 and found that factors affecting fertility are diverse. There are both personal and family factors. Personal factors include demographic factors, physical and mental health, happiness and desire to have children, and occupational status. Family factors include marital status, marital equality and satisfaction, attitudes toward gender roles, family and friend networks, and living locations and conditions. There are also macro-level factors such as cultural and social principles that influence marriage and childbearing.

Childbearing has been shown in the literature to have negative effects on mothers' labor force participation and opportunities. There are literatures on the costs of childbearing, which is measured by the wage differences between female workers with and without child. Cukrowska-Torzewska and Matysiak [7] reviewed the effects of having children on mothers' wages from 453 studies and found that women in many countries had lower wages after having children and that the average wage gap was 3.6–3.8%. Sabates-Wheeler and Kabeer [18] explains that gender inequality in most labor markets is caused by the division of labor between men and women, with women doing domestic work and men doing labor market work. The division of labor also affect the choice of jobs and career growth of women. In addition, the possibility of childbearing also can cause the discrimination in the labor market against women, especially in cases that pregnancy and child benefits are not well-developed [19].

In Thailand, Bui and Permpoonwiwat [6] examined wage inequality between male and female workers using the Labor Force Survey (LFS) data from 1996, 2006, and 2013. The results show that the gender wage gap in Thailand reduced from 14% in 1996 to % in 2013, mainly due to the higher education and skills improvement of women compared to men. Regarding family factors, Bui and Permpoonwiwat [6]'s regression using the 2013 data shows that marriage has a positive effect on wages for male workers, but a negative effect for female workers. Paweenawat and Liao [17] examined gender wage gap and parenthood wage gap for married and unmarried women using the 1985–2017 cross-sectional LFS and the 2005–2012 panel Socioeconomic Household Survey (SES) data. The fixed-effects estimation using the SES panel data for the study's most recent birth cohort (1985–1994) shows that 15.7% of motherhood penalty for married women and 33.7% penalty for unmarried women. For men, the results show smaller effects, which are 5.2% and 30.% fatherhood penalty for married and unmarried men.

Since one of the main explanations for the wage gap in marriage and childrearing is the division of labor between housework (including childcare) and labor market work, this study examines the effects of the presence of a spouse and child in the household on women's labor market opportunities. Specifically, women and men are classified into four groups including (1) single/no spouse and no child presented in the household (No spouse, no child: NS-NC), (2) single

parent (No spouse, with child: NS-WC), (3) spouse presented in the household but no child presented in the household (With spouse, no child: WS-NC) and (4) both spouse and child presented in the household (With spouse, with child: WS-WC). This study then compares the labor market opportunities, including labor force participation, likelihood of working in a high-skilled job, monthly income and working hours, among the four groups of people with different household structures.

In this study, cross-sectional data from the 2019 Socio-Economic Survey (SES) by Thailand's National Statistical Office are used for analysis. An advantage of the SES over the Labor Force Survey (LFS) is that the SES collects monthly income data for all work statuses including employers, own account workers, contributing family workers, members of producers' cooperative and employees. The LFS has more observations per year, but only collects income data for wage employees. Because the effects of the presence of a spouse and a child in the household on parents' income differs for wage earners and the self-employed [5], this study chooses the SES data to also capture the income of non-employees.

In estimating the effects of the presence of a spouse and child in the household on labor market opportunities, a crucial selection bias problem arises because women with higher qualification or a focus on career success are more likely to choose to remain single or not have children. As these women are more likely to work and earn higher income, the estimated impact can be overestimated [20]. For this reason, this study adopted the Multinomial treatment model by Deb [8] as the model can correct the selection bias. Because the effects of marriage and childbearing on labor force participation and opportunities for women are generally higher in the years of marriage and childbearing or for younger mothers [15, 17], this study focuses on examining the effects on young female adults aged 18 to 35 years old<sup>3</sup>. In addition, this study also compares the effects of the presence of a spouse and a child in the household on women's labor market opportunities with those of men.

## 2 Methodology

As the estimation of the effects of the presence of a spouse and child in the household on labor market opportunities faces the bias due to women and men's selection into living with their spouses and children, this study adopts the multinomial treatments and continuous, count and binary outcomes model or the multinomial treatment model (MTM) by Deb [8]. The MTM model estimates the treatment effects of multinomial treatments on the outcome variable by

---

<sup>3</sup> An additional reason for the scope to only study young adults is the data limitation. We can only observe parent-child relationships if the child is still living in the same household as his/her parents (See the data section for more details). As children are more likely to leave home after a certain age, the bias from the unobserved parent-child relationship increases, this study only focuses on young adults aged 18–35 years old.

using a system of two equations consisting of (1) the multinomial treatments selection equation and (2) the outcome equation. The MTM model was chosen for this study for two main reasons, namely because the treatment effects are corrected for the sample selection problem and because the models can be used with different types of outcome variables, including continuous, counting, and binary variables. In addition, although recommended, the MTM model does not require exclusion restrictions [8].

### 2.1 The Multinomial Treatments Selection Equation

For the multinomial treatments selection, each individual  $i$  can choose a household structure  $j$  from the following four choices: (1) single/no spouse and no child presented in the household (No spouse, no child: NS-NC), (2) single parent (No spouse, with child: NS-WC), (3) spouse presented in the household but no child presented in the household (With spouse, no child: WS-NC) and (4) both spouse and child presented in the household (With spouse, with child: WS-WC). From the choice, the individual will receive the unobserved utility  $V_{ij}^*$ :

$$V_{ij}^* = z_i' \alpha_j + \delta_j l_{ij} + \eta_{ij} \tag{1}$$

where  $z_i$  is the exogenous variable affecting the presence of a spouse and child in the household,  $l_{ij}$  is the unobserved variables affecting,  $V_{ij}^*$ ,  $\alpha_j$  and  $\delta_j$  are parameters for each choice  $j$  and  $\eta_{ij}$  is the error term.

Let  $\mathbf{d}_i = (d_{i1}, d_{i2}, \dots, d_{ij})$  be a vector of the dummy variable for each choice  $j$  and  $\mathbf{l}_i = (l_{i1}, l_{i2}, \dots, l_{ij})$  be a vector of unobserved variables affecting  $V_{ij}^*$ . The probability of an individual  $i$  choosing each choice is given by:

$$\Pr(\mathbf{d}_i | \mathbf{z}_i \mathbf{l}_i) = g\left(\mathbf{z}_i' \alpha_1 + \delta_1 l_{i1}, \mathbf{z}_i' \alpha_2 + \delta_2 l_{i2}, \dots, \mathbf{z}_i' \alpha_j + \delta_j l_{ij}\right) \tag{2}$$

where  $g$  is the Mixed multinomial logit (MMNL) distribution.

### 2.2 The Outcome Equation

This study examines effects of the presence of a spouse and child in the household on labor market outcomes  $y_k$  in four dimensions as follows:

The expected outcome  $y_k$  is

$$E(y_{ik} | \mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \mathbf{x}_i' \beta_k + \sum_{j=0}^J \gamma_{jk} d_{ij} + \sum_{j=0}^J \lambda_{jk} l_{ij} \tag{3}$$

where  $\mathbf{x}_i$  is the exogenous factors affecting the labor market outcomes,  $\lambda_{jk}$  are the factor-loading parameters capturing the correlation of the treatment and outcome variables through unobserved characteristics and  $\gamma_{jk}$  are the treatment effects of the household structure choice  $d_j$  on the labor market outcome  $y_k$  (Table 1).

**Table 1.** Outcome variables

	Outcome	Description	Type of variable	Distributional assumption
$y_1$	Labor force participation	= 1 if in the labor force, = 0 otherwise	Binary variable	Logistic
$y_2$	High-skill job	= 1 if work in professional or technical jobs, = 0 otherwise	Binary variable	Logistic
$y_3$	Monthly income	$\ln(\text{Monthly income})$	Continuous variable	Normal
$y_4$	Working hour	$\ln(\text{Working hour per month})$	Continuous variable	Normal

Note:  $\ln(\text{Income}) = \text{sign}(\text{Income}) \cdot \ln(\text{abs}(\text{Income}) + 1)$  is the negative logarithm function, which is the logarithm transformation for both positive and negative values [14].

To determine whether each explanatory variable significantly determines the outcome, the Bayes factor upper bound (BFUB) is used instead of the traditional p-value. As stated in Halsey [11], the BFUB is the upper bound of the odds in favor of the alternative hypothesis ( $H_a : \beta, \gamma \neq 0$ ) relative to the null hypothesis ( $H_0 : \beta, \gamma = 0$ ). In particular, the BFUB can be written as a function of the p-value (p) as follows:

$$BFUB = \frac{-1}{e \cdot p \cdot \ln(p)} \tag{4}$$

It should be noted that, the BFUBs that exceed 999 are reported as 999 in this study. The BFUB equals to 999 means that the odds in favor of the ( $H_a$ ) relative to the ( $H_0$ ) is 999 to 1, which should be considered as a strong enough evidence to support that the explanatory variable significantly determines the outcome. For the comparison purpose, the p-value of 0.01 corresponds to a BFUB of only 8.13 [3].

### 3 Data

To study the effects of having child and spouse presented in the household on labor market opportunity of workers in Thailand, this study uses work, individual and household characteristics data of male and female young adult aged 18–35 years old from the 2019 Socio-Economic Survey (SES). The 2019 SES is a comprehensive cross-sectional data by Thailand’s National Statistical Office with a total sample of 124,874 individuals, 22,599 of whom are between the age of 18–35 years old.

The limitation of the SES data is that there is no variable for the relationship between parent and child. Only the variable for the relationship to the head of household is provided. Since the parent-child relationship is required for the

estimation, only the samples where the parent-child relationship can be matched are included in this study. The matching of parent and child can be done in two cases. The first case is when the parent is the head of the household or spouse of the head of household. In this case, the child can be identified directly from the relationship to the head of household variable. The second case is when the parent is the child of the head of household. In this case, the child can be identified as the grandchildren of the head of household. The parent-child matching in the second case is complicated as the head of household may have several children. For this issue, the parent-child relationship is identified by the order of data entry in the SES survey, in which the children of each child of the head of household are entered directly after their parents. It should be noted that parents and children can only be matched if they live in the same household. If the child moves out of the household, the person is classified as having no child. For this reason, the study focuses on the effects of having a child presented in the household, rather than the effects of having children in general.

The description and basic statistics of all variables used in this study are shown in Table 2.

## 4 Results and Discussion

The MTM model uses the cross-sectional data from the 2019 SES survey to simultaneously estimate a system of two equations, the multinomial treatment and outcome equations. Therefore, the results are presented in two parts. Part 1 discusses the factors that influence the decision to live with a spouse or a child, and Part 2 examines the effects of having a spouse and a child in the household on the labor market opportunities of young female and male adults aged 18 to 35 years old.

### 4.1 Factors Affecting the Decision to Live with a Spouse or a Child

The factors affecting the decision to live with a spouse or a child is examined from the multinomial treatment equation, in which an individual can choose a household structure from the four choices including (1) single/no spouse and no child presented in the household (No spouse, no child: NS-NC), (2) single parent (No spouse, with child: NS-WC), (3) spouse presented in the household but no child presented in the household (With spouse, no child: WS-NC) and (4) both spouse and child presented in the household (With spouse, with child: WS-WC). The MTM estimation provides slightly different coefficients for the multinomial treatment equations under four different outcome equations. Because the MTM assumes a logistic distribution, the results presented in this part show the fixed effects from the multinomial logit model.

In the case of young female adults in Thailand, when considering the factors with high BFUBs, education is the main factor affecting the decision to live with a spouse, while household income and living in an urban area are the main factors affecting the decision to live with a child. Specifically, individuals with a

**Table 2.** Variable description

Variables	Description	Female			Male		
		N	mean	SD	N	mean	SD
<b>Household structures</b>							
NS-NC	No spouse, no child in household	9,630	0.387	0.487	9,744	0.570	0.495
NS-WC	No spouse, with child in household	9,630	0.116	0.320	9,744	0.053	0.224
WS-NC	With spouse, no child in household	9,630	0.165	0.371	9,744	0.139	0.346
WS-WC	With spouse, with child in household	9,630	0.332	0.471	9,744	0.238	0.426
<b>Work characteristics</b>							
LFP	Labor force participation	9,630	0.730	0.444	9,744	0.854	0.353
High skilled	Professional or technical job	7,030	0.164	0.371	8,320	0.081	0.273
Income	Monthly income	7,030	11,313	12,873	8,320	10,996	14,195
Hour	Working hours per month	4,727	194	47.30	5,499	192	49.71
<b>Individual and household characteristics</b>							
Age	Age	9,630	27.64	5.052	9,744	27.46	5.158
High school	High school degree	9,630	0.243	0.429	9,744	0.217	0.412
Vocational	Vocational degree	9,630	0.057	0.231	9,744	0.068	0.252
Higher vocational	Higher vocational degree	9,630	0.065	0.247	9,744	0.080	0.271
College	College degree or above	9,630	0.262	0.440	9,744	0.143	0.350
Urban	Living in urban area	9,630	0.576	0.494	9,744	0.555	0.497
Region: BMA	Living in Bangkok Metropolitan Area	9,630	0.073	0.260	9,744	0.070	0.254
Region: Central (No BMA)	Living in the central region outside of the Bangkok Metropolitan Area	9,630	0.320	0.467	9,744	0.330	0.470
Region: North	Living in the northern region	9,630	0.193	0.394	9,744	0.177	0.381
Region: Northeast	Living in the northeastern region	9,630	0.211	0.408	9,744	0.214	0.410
Region: South	Living in the southern region	9,630	0.203	0.402	9,744	0.210	0.407
Household size	Household size	9,630	3.779	1.772	9,744	3.626	1.792
Mother in HH	Mother presented in the household	9,630	0.561	0.496	9,744	0.624	0.484
Father in HH	Father presented in the household	9,630	0.452	0.498	9,744	0.517	0.500
High income HH	Household income per member is in the 4th or 5th quintile	9,630	0.350	0.477	9,744	0.053	0.119

Source: Calculated from SES 2019 using the sample under the scope of this study.

higher education degree are less likely to get married and live with their spouse. In addition, individuals from a high-income household (household income per household member is in the 4th and 5th quantiles) who live in an urban area are less likely to have their children presented in the household.

When considering the impact of various factors on each type of the four household structures, it was found that women with higher education (high school, vocational, higher vocational or bachelor’s degree or higher) and live in high-income households in an urban area outside of the northern region are more likely to have no spouse and child presented in the household (NS-NC group). Second, women from low-income households in a rural area in the northern or northeastern regions are more likely to be single mothers (NS-WC group). There is no strong evidence suggesting that education affects the likelihood to be a

**Table 3.** Factors determining the decision to live with spouse and have children

	(F1 mfx)	(F2 mfx)	(F3 mfx)	(F4 mfx)	(M1 mfx)	(M2 mfx)	(M3 mfx)	(M4 mfx)
	NS-NC	NS-WC	WS-NC	WS-WC	NS-NC	NS-WC	WS-NC	WS-WC
College or higher	0.307 (999.0)	0.013 (1.5)	-0.123 (999.0)	-0.198 (999.0)	0.128 (999.0)	-0.020 (3.8)	-0.053 (999.0)	-0.055 (888.2)
Higher vocational	0.099 (999.0)	0.031 (4.8)	-0.082 (999.0)	-0.048 (6.4)	0.022 (1.0)	0.017 (2.4)	-0.042 (14.2)	0.004 (2.5)
Vocational	0.165 (999.0)	0.013 (1.0)	-0.100 (999.0)	-0.078 (999.0)	0.059 (83.7)	-0.005 (1.1)	-0.064 (999.0)	0.010 (1.1)
High school	0.148 (999.0)	0.007 (1.0)	-0.093 (999.0)	-0.063 (999.0)	0.061 (999.0)	0.006 (1.0)	-0.048 (999.0)	-0.019 (2.0)
Age	-0.139 (999.0)	-0.007 (1.0)	0.039 (999.0)	0.107 (999.0)	-0.170 (999.0)	0.014 (4.6)	0.052 (999.0)	0.104 (999.0)
Age2	0.002 (999.0)	0.000 (1.3)	-0.001 (434.5)	-0.002 (999.0)	0.003 (999.0)	0.000 (2.2)	-0.001 (999.0)	-0.001 (999.0)
Urban	0.038 (999.0)	-0.019 (14.9)	0.011 (1.2)	-0.030 (68.1)	0.008 (1.0)	-0.013 (10.8)	0.012 (1.6)	-0.008 (1.0)
Region: Central (Not BMA)	-0.016 (1.0)	0.016 (1.0)	0.011 (1.0)	-0.011 (1.2)	-0.035 (1.7)	0.009 (1.1)	0.014 (1.0)	0.012 (1.1)
Region: North	-0.032 (1.6)	0.039 (3.1)	-0.071 (999.0)	0.064 (15.7)	-0.036 (1.5)	0.011 (1.0)	-0.045 (20.8)	0.071 (51.0)
Region: Northeast	-0.025 (1.2)	0.080 (999.0)	-0.092 (999.0)	0.038 (1.7)	0.026 (1.1)	0.028 (2.8)	-0.081 (999.0)	0.026 (1.1)
Region: South	0.012 (1.1)	-0.011 (1.2)	-0.052 (68.1)	0.050 (4.1)	-0.005 (2.2)	-0.004 (1.8)	-0.040 (11.0)	0.050 (4.6)
High income HH	0.163 (999.0)	-0.097 (999.0)	0.103 (999.0)	-0.169 (999.0)	0.162 (999.0)	-0.059 (999.0)	0.064 (999.0)	-0.167 (999.0)
Observation	9,630				9,744			
	AIC = 20962, BIC = 21242, McFadden's R2 = 0.153, McFadden's Adjusted R2 = 0.149				AIC = 18853, BIC = 19133, McFadden's R2 = 0.118, McFadden's Adjusted R2 = 0.114			

*Note:* (1) Bayes factor upper bound (BFUB) in parentheses. (For BFUB higher than 999, the value 999 is reported.) (2) Results shown are marginal effects from the multinomial Logit regression. (3) The education variables consist of 4 dummy variables with the base group being junior high school or lower. (4) The region variables consist of 4 dummy variables with the base group being the Bangkok and Metropolitan Area (BMA).

single mother, except in the case of higher vocational education, which increases the likelihood of being a single mother.

Third, women with low education who live in high-income households in the Bangkok metropolitan area or other provinces in the central region are most likely to live with a husband but have no children (WS-NC group). Finally, women with low education who live in low-income households in a rural area in the north, northeast or south of Thailand are likely to live with a husband and have children (WS-WC group).

For young male adults, factors affecting the decision to live with a spouse and a child does not differ from those affecting women. That is, education has the greatest impact on the decision to live with a spouse, while household income and living in an urban area have the greatest impact on the decision to live with their children. However, the sizes of the effects are smaller than those of women.



## 4.2 The Effects of the Presence of a Spouse and Child in the Household on Labor Market Opportunities

This section compares the labor market opportunities of male and female young adults in four dimensions including (1) labor force participation (LFP), (2) working in a high skill job, (3) monthly income, and (4) working hours using the SES 2019 data. For the sample used in this study, female young adults were less likely to be in the labor force compared to male. However, among those who were employed, a higher proportion of females held higher-skilled jobs (professional and technical occupations) and earned higher average incomes compared to males. This could be the effect of education. While there was a higher proportion of male young adults with vocational and higher vocational degrees, there was a higher proportion of female young adults with college degrees (See Table 2). It should be noted that, for the overall working age population, male workers still earned higher monthly income than female workers. For working hours, there were no major differences between male and female workers in terms of working hours.

**Table 4.** Labor market statistics for male and female young adults

Group	Female				Male			
	LFP	High skilled	Income	Hour	LFP	High skilled	Income	Hour
1. NS-NC	68.07%	27.04%	12,810	193	76.94%	10.12%	10,240	189
2. NS-WC	74.44%	13.01%	10,153	191	88.22%	4.38%	8,620	183
3. WS-NC	85.60%	10.51%	11,662	205	97.20%	6.82%	12,177	205
4. WS-WC	71.98%	9.51%	9,875	186	98.06%	5.76%	12,207	189
Total	73.00%	16.44%	11,313	193	85.39%	8.09%	10,996	192

Source: Calculated from SES 2019 using the sample under the scope of this study (described in the data section).

For the MTM regressions, the results in this section illustrate the effects of the presence of a spouse and child on (1) labor force participation (LFP), (2) working in a high skill job, (3) monthly income, and (4) working hours of Thai young adults. The model compensates for the bias caused by the selection into living with a spouse and a child using the variables listed in the selection equation (see Table 3).

This study provides estimates of the effects in three models- Model (a), (b) and (c)- with a different set of independent variables as shown in Table 5. The results show a small difference in the estimates of the effects in all four dimensions indicating that the estimations are quite robust to the choice of controlled variables. Moreover, in Model (b), which includes the urban and region dummies as independent variables, shows that not all urban and region dummies are statistically significant in most of the outcome models. As the urban and region dummies are included in the selection equation of all models, the variables may act as exclusion restrictions in these models<sup>4</sup>.

<sup>4</sup> Although recommended, the MTM model does not require exclusion restrictions [8].

For the results, the coefficients measuring the effects of having a spouse and a child in the household shown in Table 5 and 6 are to be compared with the base group, which is the no spouse- no child (NS-NC) group. Overall, the results show that, controlling for education, age and the selection bias, there are strong evidence suggesting that the presence of a spouse and child in the household affects both men and women’s labor force participation, monthly income, and

**Table 5.** Factors affecting the economic participation and opportunity of female young adults (age 18–35).

	LFP			High skill job		
	F1(a)	F1(b)	F1(c)	F2(a)	F2(b)	F2(c)
<b>Effects of presence of spouse or child in the household</b>						
(2) NS-WC	0.03 (999.0)	-0.09 (1.5)	0.02 (1.3)	0.00 (1.0)	-0.01 (1.0)	-0.01 (1.0)
(3) WS-NC	0.13 (59.4)	0.14 (999.0)	0.13 (135.1)	0.00 (1.3)	0.00 (1.6)	0.00 (1.3)
(4) WS-WC	-0.27 (11.0)	-0.26 (43.4)	-0.28 (15.4)	-0.01 (1.0)	-0.01 (1.0)	-0.01 (1.0)
<b>Effects of individual and household characteristics</b>						
College or higher	0.09 (482.1)	0.10 (999.0)	0.09 (314.1)	0.44 (999.0)	0.45 (999.0)	0.44 (999.0)
Higher vocational	0.00 (2.7)	0.00 (55.8)	-0.01 (2.0)	0.05 (1.1)	0.05 (1.1)	0.05 (1.1)
Vocational	0.08 (729.8)	0.08 (999.0)	0.07 (434.5)	0.22 (7.6)	0.24 (11.9)	0.23 (8.1)
High school	-0.03 (1.5)	-0.02 (1.5)	-0.03 (1.7)	0.01 (1.0)	0.01 (1.0)	0.01 (1.0)
Age	0.21 (999.0)	0.20 (999.0)	0.21 (999.0)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Age2	0.00 (999.0)	0.00 (999.0)	0.00 (999.0)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Urban		-0.03 (30.5)			0.00 (1.6)	
Region: Central (not BMA)		0.05 (8.8)			0.00 (1.0)	
Region: North		0.04 (1.9)			0.00 (1.0)	
Region: Northeast		0.08 (999.0)			0.00 (1.0)	
Region: South		0.00 (3.2)			0.00 (1.0)	
Father in HH			-0.02 (2.3)			0.00 (2.0)
Mother in HH			0.04 (14.5)			0.00 (1.2)
Constant						
Observations	9,630	9,630	9,630	7,030	7,030	7,030
AIC	30,313	30,270	30,309	20,304	20,298	20,308
BIC	30,686	30,679	30,696	20,661	20,688	20,678

(continued)

Table 5. (continued)

	nln(Income)			ln(Hour)		
	F3(a)	F3(b)	F3(c)	F4(a)	F4(b)	F4(c)
<b>Effects of presence of spouse or child in the household</b>						
(2) NS-WC	-0.83 (10.5)	-0.69 (1.5)	-0.84 (165.3)	-0.15 (999.0)	-0.12 (108.0)	-0.16 (999.0)
(3) WS-NC	0.81 (999.0)	0.47 (2.6)	0.61 (14.8)	0.00 (9.1)	-0.03 (1.0)	0.00 (8.4)
(4) WS-WC	-0.98 (32.8)	-0.64 (1.3)	-0.64 (3.7)	-0.21 (999.0)	-0.17 (999.0)	-0.21 (999.0)
<b>Effects of individual and household characteristics</b>						
College or higher	0.76 (999.0)	0.78 (999.0)	1.00 (999.0)	0.66 (999.0)	0.67 (999.0)	0.65 (999.0)
Higher vocational	0.50 (12.2)	0.47 (8.8)	0.63 (110.1)	0.35 (999.0)	0.34 (999.0)	0.34 (999.0)
Vocational	0.65 (999.0)	0.65 (888.2)	0.79 (999.0)	0.32 (999.0)	0.33 (999.0)	0.32 (999.0)
High school	-0.03 (2.0)	0.00 (14.3)	0.05 (1.2)	0.20 (999.0)	0.20 (999.0)	0.20 (999.0)
Age	0.39 (224.9)	0.36 (68.1)	0.32 (26.7)	0.05 (2.2)	0.05 (2.4)	0.05 (2.4)
Age2	-0.01 (56.1)	-0.01 (25.5)	-0.01 (15.6)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Urban		-0.06 (1.0)			-0.02 (1.0)	
Region: Central (not BMA)		-0.38 (6.8)			-0.18 (999.0)	
Region: North		-0.85 (999.0)			-0.34 (999.0)	
Region: Northeast		-1.43 (999.0)			-0.28 (999.0)	
Region: South		-0.48 (15.9)			-0.28 (999.0)	
Father in HH			-0.20 (3.4)			0.02 (1.0)
Mother in HH			-0.58 (999.0)			0.00 (15.1)
Constant	2.08 (1.3)	3.26 (4.0)	3.68 (8.4)	2.78 (999.0)	3.00 (999.0)	2.75 (999.0)
Observations	7,030	7,030	7,030	4,687	4,687	4,687
AIC	51,674	51,549	51,598	18,951	18,854	18,954
BIC	52,038	51,947	51,976	19,293	19,229	19,309

Note: (1) Bayes factor upper bound (BFUB) in parentheses. (For BFUB higher than 999, the value 999 is reported.) (2) The results were estimated using the multinomial treatments model, which adjusted for the selection into living with spouse and having a child. (3) nln(Income) and ln(Hour) was calculated using negative log function, which is  $nln(X) = sign(X) * ln(abs(X)+1)$ . (4) Working hour is only observed for government, state enterprise and private employees. (5) The skill position, wage and working hour estimations were not corrected for the selection into the labor force and, thus, the results should be interpreted for the employed population (or employees for the working hour regression) rather than the working-age population. (6) The high skilled position is defined as occupations with professional and technical skills (Skill class 2 and 3 in the ISCO-08 classification).

working hours. However, the household structure does not directly affect an opportunity to work in a high skill job for both men and women. From Model F2 and M2 in Table 5 and 6, only a higher educational degree significantly improves the opportunity to work in a high skill job.

Consider the MTM results for labor force participation, monthly income and working hours in the case of women in Table 5 (Model F1(c), F3(c) and

**Table 6.** Factors affecting the economic participation and opportunity of male young adults (age 18–35).

	LFP			High skill job		
	M1(a)	M1(b)	M1(c)	M2(a)	M2(b)	M2(c)
<b>Effects of presence of spouse or child in the household</b>						
(2) NS-WC	0.00 (1.9)	0.00 (2.9)	0.01 (1.1)	0.00 (1.1)	0.00 (1.6)	0.00 (1.1)
(3) WS-NC	0.07 (999.0)	0.07 (999.0)	0.07 (999.0)	0.00 (1.1)	0.00 (1.0)	0.00 (1.3)
(4) WS-WC	0.07 (542.4)	0.07 (999.0)	0.07 (235.9)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
<b>Effects of individual and household characteristics</b>						
College or higher	0.00 (1.2)	0.00 (1.3)	0.00 (1.2)	0.22 (235.9)	0.23 (621.4)	0.22 (155.3)
Higher vocational	-0.09 (999.0)	-0.09 (999.0)	-0.09 (999.0)	0.00 (1.0)	0.01 (1.0)	0.00 (1.0)
Vocational	0.03 (58.8)	0.03 (190.2)	0.03 (68.9)	0.03 (1.1)	0.04 (1.4)	0.02 (1.1)
High school	-0.05 (999.0)	-0.04 (999.0)	-0.05 (999.0)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Age	0.07 (999.0)	0.07 (999.0)	0.07 (999.0)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Age2	0.00 (999.0)	0.00 (999.0)	0.00 (999.0)	0.00 (1.0)	0.00 (1.0)	0.00 (1.0)
Urban		-0.03 (999.0)			0.00 (1.0)	
Region: Central (not BMA)		0.01 (1.7)			0.00 (1.0)	
Region: North		0.00 (1.7)			0.00 (1.0)	
Region: Northeast		0.00 (2.3)			0.00 (1.0)	
Region: South		0.02 (5.5)			0.00 (1.0)	
Father in HH			0.01 (1.0)			0.00 (1.1)
Mother in HH			-0.02 (47.9)			0.00 (1.0)
Constant						
Observations	9,744	9,744	9,744	8,320	8,320	8,320
AIC	24,651	24,607	24,644	20,689	20,666	20,690
BIC	25,025	25,016	25,032	21,054	21,066	21,069

(continued)

Table 6. (continued)

	nln(Income)			ln(Hour)		
	M3(a)	M3(b)	M3(c)	M4(a)	M4(b)	M4(c)
<b>Effects of presence of spouse or child in the household</b>						
(2) NS-WC	-0.27 (1.2)	-0.11 (1.2)	0.07 (1.7)	-0.11 (8.1)	-0.07 (1.8)	-0.09 (2.9)
(3) WS-NC	0.92 (999.0)	0.63 (888.2)	0.50 (70.5)	0.11 (294.3)	0.05 (1.6)	0.10 (43.4)
(4) WS-WC	-0.81 (999.0)	-0.74 (999.0)	-0.80 (999.0)	-0.11 (999.0)	-0.06 (6.8)	-0.10 (729.8)
<b>Effects of individual and household characteristics</b>						
College or higher	0.49 (999.0)	0.47 (999.0)	0.58 (999.0)	0.62 (999.0)	0.61 (999.0)	0.63 (999.0)
Higher vocational	0.37 (6.9)	0.33 (4.0)	0.43 (19.5)	0.23 (999.0)	0.23 (999.0)	0.24 (999.0)
Vocational	0.55 (999.0)	0.61 (999.0)	0.65 (999.0)	0.35 (999.0)	0.37 (999.0)	0.36 (999.0)
High school	-0.17 (2.2)	-0.09 (1.0)	-0.10 (1.1)	0.17 (999.0)	0.19 (999.0)	0.17 (999.0)
Age	0.46 (999.0)	0.45 (999.0)	0.38 (999.0)	0.07 (66.6)	0.07 (125.5)	0.07 (46.2)
Age2	-0.01 (999.0)	-0.01 (999.0)	-0.01 (888.2)	0.00 (6.0)	0.00 (8.5)	0.00 (5.2)
Urban		-0.25 (160.1)			0.01 (1.2)	
Region: Central (not BMA)		-0.52 (165.3)			-0.20 (999.0)	
Region: North		-1.08 (999.0)			-0.36 (999.0)	
Region: Northeast		-1.76 (999.0)			-0.44 (999.0)	
Region: South		-0.70 (999.0)			-0.27 (999.0)	
Father in HH			-0.47 (999.0)			0.01 (1.9)
Mother in HH			-0.78 (999.0)			-0.05 (7.6)
Constant	1.02 (1.0)	2.03 (2.0)	3.09 (12.6)	2.68 (999.0)	-16.10 (999.0)	2.75 (999.0)
Observations	8,320	8,320	8,320	5,392	5,392	5,392
AIC	59,169	58,930	58,924	20,149	19,922	20,144
BIC	59,541	59,338	59,310	20,498	20,304	20,506

Note: (1) Bayes factor upper bound (BFUB) in parentheses. (For BFUB higher than 999, the value 999 is reported.) (2) The results were estimated using the multinomial treatments model, which adjusted for the selection into living with spouse and having a child. (3) nln(Income) and ln(Hour) was calculated using negative log function, which is  $nln(X) = sign(X) * ln(abs(X)+1)$ . (4) Working hour is only observed for government, state enterprise and private employees. (5) The skill position, wage and working hour estimations were not corrected for the selection into the labor force and, thus, the results should be interpreted for the employed population (or employees for the working hour regression) rather than the working-age population. (6) The high skilled position is defined as occupations with professional and technical skills (Skill class 2 and 3 in the ISCO-08 classification).

F4(c))<sup>5</sup>. After controlling for educational level, age and the presence of parents in the household and correcting for the selection to living with a spouse or a child, single mothers (NS-WC) are not more likely to enter the labor market compared with the base group (NS-NC). Female young adults living with their spouse but without child (WS-NC) are 13% more likely and women living with their spouse and child (WS-WC) are 28% less likely to be in the labor market. It can be noticed that the effects of the household structure on labor force participation estimated using the MTM model is smaller than suggested by the raw data calculation. From Table 3, all three groups- NS-WC, WS-NC and WS-WC- participate in the labor market more than the NS-NC group by 6.37%, 17.53% and 3.91%, respectively. This shows that an estimate without controlled variables or sample selection correction would be upward bias.

Regarding monthly income and working hours, female young adults who live with their spouse but do not have a child (WS-NC) earn 61% higher monthly income compared to the base group (NS-NC), but do not have evidence to support that they have higher working hours. That is, living with a spouse without a child influences women to participate more in the labor force and also increase female workers' income. For female workers who have a child in the household (NS-WC and WS-WC), they work significantly fewer hours and earn significantly lower monthly income. Female workers with a child and no spouse in the household (NS-WC) face 84% lower income and female workers with a spouse and a child (WS-WC) face 64% lower income. The results suggest that the division of labor, in which women do more household work and men do more labor market work, only affect women's participation and opportunities in the labor market only when they have a child.

The child effect results are partly consistent with Paweenawat and Liao [17]'s finding that unmarried women face a higher parenthood penalty compared to married women. However, the size of the income gap is significantly higher in this study. This is potentially because Paweenawat and Liao [17] analyze parenthood penalty using hourly wage gap, while this study uses monthly income gap and the monthly income contains the working hour effect. In particular, female workers with children are likely to work fewer hours and, thus, reduce their monthly income but not their hourly wage. Moreover, this study covers all types of employment status and not just employees. As the parenthood penalty is potentially higher for self-employed workers than for wage employees [5], the effects in this study are expected to be higher. In addition to the differences due to the choice of measurement and the scope of the study, there is also a difference in the estimation methods to control for endogeneity bias due to ability or other labor market qualifications of women who have and do not have children. While Paweenawat and Liao [17] uses fixed effects for the 2005–2012 SES panel data,

<sup>5</sup> Although Models F1(b), F2(b), F3(b) and F4(b) with regions as controlled variables in the outcome equations have lower AIC and BIC, Models F1(c), F2(c), F3(c) and F4(c) are chosen for the analysis for the purpose of leaving region variables as exclusion restrictions. It should be noted that there is no sign difference in the estimates across the models and the effect sizes only vary slightly.

this study uses 2019 cross-sectional data with the MTM model to correct for the selection to have and live with children (or spouse)<sup>6</sup>. With this cross-sectional limitation, the MTM model only can account for the selection bias and not all other ability bias, which can cause an upward bias.

Effects of the presence of a spouse and child in the household on labor force participation, monthly income and working hours for male are shown in Table 6 (Model M1(c), M2(c) and M4(c))<sup>7</sup>. For men, single fatherhood (NS-WC) does not affect any dimensions of participation and opportunity in the labor market, except that they work slightly fewer hours comparing to the base group (NS-NC). Similar to the case for women, male workers living with spouse but no children (WS-NC) tend to participate in the labor force more and earn higher income compared to the base group (NS-NC). However, in the case of WS-WC, the results are different between men and women. While female workers living with a spouse and child are less likely to participate in the labor market, male workers are more likely to work. Both male and female workers living with a spouse and child have lower working hours and lower incomes.

For other factors affecting participation and opportunities in the labor market, education is the most important factor. Higher education leads women to work more in the labor market, but has less impact on men's labor market entry, as Thai men have a higher labor force participation rate regardless of education level. In addition to the labor force participation dimension, higher education provides both male and female workers with the opportunity to learn a highly skilled job and increase their income and working hours.

For family factors, living with the father or mother has different effects on male and female labor force participation. Living with the father has no effect on men's labor market entry, but leads to lower participation among women. Living with the mother allows more women to enter the labor market but leads to lower labor force participation for men. There is no evidence suggesting that living with a parent has an effect on hours worked, but there are quite strong evidence that it causes both women and men to have lower monthly incomes.

## 5 Conclusion

This study examines the effects of the presence of a spouse and child in the household on women's labor market opportunities, including labor force participation, likelihood of working in a high-skilled job, monthly income and working hours. The results of the basic statistics (Table 4) show that women who are single or do not live with their spouses and have no children are more likely to participate in the labor force, work in high-skilled positions and have high income. In contrast,

<sup>6</sup> To the best of authors' knowledge, there is no SES panel data in recent years. Without the panel data, the fixed effects model cannot be estimated.

<sup>7</sup> Similar to the case of female regressions, Models M1(c), M2(c), M3(c) and M4(c) are chosen for the analysis for the purpose of leaving region variables as exclusion restrictions. There is also no sign difference in the estimates across the models and the effect sizes only vary slightly.

men who live with their spouses are more likely to participate and have a higher opportunity in the labor market.

Adopting the Multinomial Treatment Model (MTM) to correct for the selection to live with a spouse and a child, the results suggest that the division of labor, in which women do more work in the household and men do more work in the labor market, only affects women's labor market participation and opportunities if they have a child. The presence of a spouse in the household not only does not reduce the likelihood of female young adults entering the labor market or reduce their incomes, but also leads them to work more and earn higher incomes. However, the presence of children reduces the likelihood that women will enter the labor market, reduce their hours of work, and reduce their incomes. For the opportunity to work in a high skill job, only education is found to be a significant factor. That is, for both male and female young adults, the household structure does not directly affect an opportunity to work in a high skill job. For men, the effects of the presence of a spouse and child in the household are similar to those for women, except for the labor force participation dimension in the WS-WC case. Specifically, while female workers living with a spouse and a child are less likely to participate in the labor market, male workers are more likely to work.

The difference between the MTM results and basic statistics highlights the importance of including controlled variables and correcting for selection bias when estimating the effects of household structure. This indicates that the differences in the participation and economic opportunities for women with different family structure are partially caused by third factors and the selection to live with a spouse or child. This study found that the most important factor is education. Women with children tend to have lower education and, thus, face a lower opportunity in the labor market. Consequently, the key policy recommendation from this study is that, regardless of family structure, it is important to build human capital through education for female adolescents to enhance skills and improve opportunities in the labor market.

It should be noted that the MTM only corrects for selection bias and cannot fully control for ability bias. In addition, due to data limitations, this study focuses only on the short-term effects of household structure on labor market participation and opportunities for young adults. For young adults, earlier labor market entry and higher earnings do not necessarily guarantee better labor market outcomes later in life. This is mainly because those who enter the labor market early are also more likely to leave the school system early. Therefore, it is necessary to also examine the longer-term or lifetime effects.

**Acknowledgment.** This study was supported by National Research Council of Thailand (NRCT) under the Khonthai 4.0 spearhead program.



## References

1. Becker, G.S.: A theory of marriage: Part I. *J. Polit. Econ.* **81**(4), 813–846 (1973)
2. Becker, G.S.: A theory of marriage: Part II. *J. Polit. Econ.* **82**(2), S11–S26 (1974)
3. Benjamin, D.J., Berger, J.O.: Three recommendations for improving the use of p-values. *Am. Stat.* **73**(sup1), 186–191 (2019)
4. Browning, M., Chiappori, P.A., Weiss, Y.: *Economics of the Family*. Cambridge University Press, Cambridge (2014)
5. Budig, M.J.: Gender, self-employment, and earnings: the interlocking structures of family and professional status. *Gender Soc.* **20**(6), 725–753 (2006)
6. Bui, M.T.T., Permpoonwiwat, C.K.: Gender wage inequality in Thailand: a sectoral perspective. *J. Behav. Sci.* **10**(2), 19–36 (2015)
7. Cukrowska-Torzewska, E., Matysiak, A.: The motherhood wage penalty: a meta-analysis. *Soc. Sci. Res.* **88**, 102416 (2020)
8. Deb, P.: MTREATREG: Stata module to fits models with multinomial treatments and continuous, count and binary outcomes using maximum simulated likelihood. *Stat. Softw. Components* (2009)
9. Deb, P., Trivedi, P.K.: Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. *Stand. Genomic Sci.* **6**(2), 246–255 (2006)
10. Duvander, A.Z., Kridahl, L.: Decisions on marriage? Couples' decisions on union transition in Sweden. *Genus* **76**(1), 1–21 (2020)
11. Halsey, L.G.: The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* **15**(5), 20190174 (2019)
12. Hashemzadeh, M., Shariati, M., Mohammad Nazari, A., Keramat, A.: Childbearing intention and its associated factors: a systematic review. *Nurs. Open* (2021)
13. ILO. UN Women & Women Count. The Impact of Marriage and Children on Labour Market Participation (2020). [https://sustainabledevelopment.un.org/content/documents/2701theimpactofmarriageandchildrenonlabourmarketparticipation\\_en.pdf](https://sustainabledevelopment.un.org/content/documents/2701theimpactofmarriageandchildrenonlabourmarketparticipation_en.pdf)
14. John, J.A., Draper, N.R.: An alternative family of transformations. *J. Roy. Stat. Soc.: Ser. C (Appl. Stat.)* **29**(2), 190–197 (1980)
15. Loughran, D.S., Zissimopoulos, J.M.: Why wait? The effect of marriage and child-bearing on the wages of men and women. *J. Hum. Resour.* **44**(2), 326–349 (2009)
16. Mincer, J.A.: *Schooling, Experience, and Earnings*. NBER Press (1974)
17. Paweenawat, S.W., Liao, L.: Parenthood penalty and gender wage gap: recent evidence from Thailand. *J. Asian Econ.* **78**, 101435 (2022)
18. Sabates-Wheeler, R., Kabeer, N.: Gender equality and the extension of social protection (2005)
19. Waldfogel, J.: Understanding the 'family gap' in pay for women with children. *J. Econ. Perspect.* **12**(1), 137–156 (1998)
20. Wilde, E.T., Batchelder, L., Ellwood, D.T.: The mommy track divides: the impact of childbearing on wages of women of differing skill levels (No. w16582). National Bureau of Economic Research (2010)



# Link Between Renewable and Non-renewable Energy Consumption and Co2 Emissions: A Monte-Carlo Simulation Study

Phan Thi Lieu<sup>1,2</sup>(✉) and Nguyen Ngoc Thach<sup>3</sup>

<sup>1</sup> University of Labour and Social Affairs, 1018 To Ky Street,  
District 12, Ho Chi Minh City, Vietnam  
lieupt@ldxh.edu.vn

<sup>2</sup> University of Economics and Law, Vietnam National University, 669 National Highway 1,  
Linh Xuan Ward, Thu Duc, Ho Chi Minh City, Vietnam

<sup>3</sup> Banking University, 36 Ton That Dam Street,  
District 1, Ho Chi Minh City, Vietnam  
thachnn@buh.edu.vn

**Abstract.** Energy consumption is an indispensable need in economic activities as well as in daily life. However, excessive use of energy sources, particularly nonrenewable energy, leads to the depletion of natural resources and significant environmental impacts. Most of previous studies, where the link between energy use and environmental pollution is analyzed within the frequentist framework, could not meaningfully interpret p-value non-significance. In the present research, by using the generalized least squares (GLS) frequentist method in comparison with a Bayesian inference, the authors clarify the effect of each type of energy on CO<sub>2</sub> emissions in 5 ASEAN countries during the period 1990–2019. According to the outcomes obtained both by the Bayesian and frequentist methods, non-renewable energy consumption exacerbates environmental issues by increasing CO<sub>2</sub> emissions.

**Keywords:** Frequentist · Bayesian · renewable energy · non-renewable energy · CO<sub>2</sub> emissions

## 1 Introduction

Economic growth and environmental degradation are two of the three pillars of sustainable development: economic development (especially economic growth), social development (especially the realization of social progress and justice; hunger eradication and poverty reduction and job creation) and environmental protection (treatment and remedy pollution recovery, environmental restoration and improvement; fire prevention and deforestation; rational exploitation and economical use of natural resources). Nevertheless, achieving both economic growth and environmental protection at the same time is not easy. Economic growth is closely related to energy consumption because accelerated economic development necessitates increased energy consumption. (Halicioglu, 2009; Ozturk & Acaravci, 2010). In turn, energy consumption is regarded as the primary cause

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 376–383, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_26](https://doi.org/10.1007/978-3-031-35763-3_26)

of environmental problems. CO<sub>2</sub> emissions from energy consumption have significantly increased in newly-industrialized countries since the 1990s compared to industrialized countries. The deterioration of environmental quality has reached alarming levels, raising concerns about global warming and climate change (Kasman & Duman, 2015).

With the rapid development of the Southeast Asian economies in recent years, the region's energy demand has been rising fast. As a consequence, total primary energy supply could increase up to 250%, resulting in 10 to 100% more emissions by 2040 (Umbach, 2021). Southeast Asia is one of the few regions of the world where coal-fired generation has been expanding, with close to 20 GW of new coal-fired generating capacity under construction, mostly in Indonesia (a major coal producer), Viet Nam and the Philippines (IEA, 2020). The ASEAN primary energy mix is still based on fossil fuels, with a regional electricity demand increasing by 6% annually, one of the highest increases in the world (Umbach, 2021). Faced with that situation, ASEAN had an Action Plan on Energy Cooperation (APAEC) for 2010–2025 that includes short-term targets for a 23% share of renewable energy in the total primary energy supply. By 2025, ASEAN nations hope to have 35% of their energy consumption from renewable sources. Solar PV capacity needs to increase from 32 gigawatts (GW) to 83 GW and capacity of hydropower from 59 GW in 2020 to 77 GW in 2025. As ASEAN becomes a net importer of fossil fuels over the next decade, the bloc's net energy trade deficit could reach more than \$300 billion a year (Umbach, 2021).

Deriving from a diversity of studies on the same topic and the practical context of the researched countries, this article focuses on analyzing the role of renewable and non-renewable energy consumption on CO<sub>2</sub> emissions of the main ASEAN countries. The novelty of the current work compared to other studies is to examine the role of renewable and non-renewable energy in CO<sub>2</sub> emissions through both frequentist and Bayesian inference.

## 2 Literature Review

### 2.1 Key Concepts and Theories

According to Hersh (2006), renewable energy is defined as energy flows which are continuously replenished by natural processes. Renewable energy sources are divided into three main categories: (1) Direct uses of solar energy: solar thermal energy conversion and solar photovoltaics; (2) Indirect uses of solar energy: hydropower, wind power, wave power and bioenergy or biofuels; (3) Sources of renewable energy which do not depend on solar radiation: tidal and geothermal energy.

A non-renewable resource refers to a natural resource that is found beneath the earth, which when consumed, does not replenish at the same speed at which it is used up. The resources typically take millions of years to develop. The main examples of non-renewable resources are fuels such as oil, coal, and natural gas, which humans regularly exploit to produce energy. The two broad categories of non-renewable resources are fossil fuels and nuclear energy (from uranium ore).

Environmental Kuznets curve (EKC) theory should be analyzed in our study. Grossman and Krueger (1991) asserted in his study that emissions increase with an increase in per capita income for low-income countries, but decrease in countries which have higher

levels of income. In other words, they argue that there is an inverted U-shaped relationship between GDP growth and pollutant emission. This result is similar to the well-known Kuznets Curve shape in inequality research. Since then, this inverted U-shaped curve also soon became known as the EKC.

Since the EKC hypothesis was developed, many studies have been carried out to test its validity. These studies can be divided into three groups: (1) The first study group first estimated the fundamental relationship between economic output and economic output squared with environmental pollution; (2) The second group of researchers expanded on the basic relationship, adding an element of energy consumption; and (3) The third group includes additional independent variables such as urbanization, trade openness, foreign direct investment, financial development, and tourism (Verbic et al., 2021).

## 2.2 Empirics

By utilizing the OLS method and Vector Error Correction Model (VECM) on the data for the period 1982–2017, Abbasi et al. (2020) explored the impact of energy consumption, economic growth on CO<sub>2</sub> emissions in the ASEAN countries (excluding Vietnam). The result shows that when economic growth increased by 1% and energy consumption increased by 1%, CO<sub>2</sub> emissions increased by 0.4% and 1.55%, respectively. This confirms that increased energy use would exacerbate environmental degradation in the ASEAN countries. Kasman and Duman (2015) conducted a study on the impact of energy consumption on CO<sub>2</sub> emissions in EU countries for the 1992–2010 period by FMOLS estimation method and the results showed that a 1% increase in per capita energy consumption tends to increase per capita emissions by 0.56%. These assertions are consistent with Abbasi et al. (2020); Adedoyin and Zakari (2020); Kotroni et al. (2020); Sun et al. (2020); Wasti and Zaidi (2020); Dehghan Shabani and Shahnazi (2019).

Salari et al. (2021) performed an analysis of the relation between renewable energy and non-renewable energy and CO<sub>2</sub> emissions in the US during 1997–2016 employing Ordinary Least Squared (OLS) and Generalized method of moments (GMM) methods. A 47% increase in energy use leads to a 31% corresponding increase in CO<sub>2</sub> emissions. More specifically, renewable energy tend to reduce CO<sub>2</sub> emissions while non-renewable energy to increase this indicator. Similar finding are also found in Bekun et al. (2019).

In sum, all earlier works considered in our study used outdated frequentist estimators.

## 3 Research Model, Data Sources, Method

The article focuses on analyzing effect of renewable and non-renewable energy consumption on carbon emissions in 5 main ASEAN countries. Based on the theoretical background and previous empirical studies, we propose a general econometric model as follows:

$$CO_{2it} = \beta_0 + \beta_1 REC_{it} + \beta_2 EC_{it} + \beta_3 Gr_{it} + \beta_4 PoP_{it} + u_{it}$$

where  $u$  is the error term,  $t$  is the years (from 1990 to 2020);  $i$  is the country (Malaysia, Philippines, Singapore, Thailand, Vietnam).

To achieve the research objectives, the impact of renewable and non-renewable energy on CO<sub>2</sub> emissions is considered separately. Population is one of the impacts on environmental quality, according to the IPAT and STIRPAT theoretical models, so the authors included the population growth rate variable in the study model. In addition, the authors also inherited the study of Abbasi et al. (2020) on the proposal to turn economic growth to CO<sub>2</sub> emissions (Table 1).

**Table 1.** Information about variables, units and data sources

Variable	Variable explanation	Unit	Sources
CO <sub>2</sub>	Annual CO <sub>2</sub> emissions per unit energy	kg per kilowatt-hour	Our World in Data
Gr	GDP growth (annual)	%	World Bank
PoP	Population growth (annual)	%	World Bank
REC	Renewables per capita	kWh	Our World in Data
EC	Fossil fuels per capita	kWh	Our World in Data

*Sources: Author*

Variables REC, EC are logarithmic to smooth the data.

For comparison purpose, the GLS frequentist and Bayesian regressions are used to examine the role of renewable and non-renewable energy in CO<sub>2</sub> emissions. We employ the Metropolis-Hasting and Gibbs samplers in comparison with the Pool-OLS, Fixed Effects (FEM), and Random Effects (REM), which are fixed by Generalized least squares (GLS) method. Within the Bayesian approach, the informative normal priors (0,1) are assigned to all the model parameters as recommended by Block et al. (2011). Let's mention that as our sample size is large, different prior specifications have small effect on the posterior distribution.

## 4 Empirical Results

### 4.1 Frequentist vs. Bayesian Outcomes

As mentioned above, in this article, the authors use frequentist estimators (Pool-OLS, FEM, REM, GLS) and Bayes-based MCMC samplers for comparison. Among the frequentist methods, model selection and other necessary tests are carried out to detect heteroscedasticity and autocorrelation. Within the frequentist framework, the model is corrected by GLS method whose results are presented in Table 2 (column 4), whereas the Bayesian ones are demonstrated in Table 3.

**Table 2.** Frequentist estimation outcomes

	(1)	(2)	(3)	(4)
	CO <sub>2</sub>	CO <sub>2</sub>	CO <sub>2</sub>	CO <sub>2</sub>
REC	-0.0426 [-0.32]	-0.143 [-0.54]	-0.0426 [-0.32]	0.0144 [0.15]
EC	2.091*** [37.60]	1.396*** [4.68]	2.091*** [37.60]	1.650*** [16.25]
PoP	0.386*** [4.16]	-0.0358 [-0.34]	0.386*** [4.16]	0.0492 [0.84]
Gr	0.0533** [2.13]	0.00673 [0.30]	0.0533** [2.13]	0.00841* [1.77]
_cons	-16.77*** [-30.72]	-9.282*** [-3.09]	-16.77*** [-30.72]	-12.32*** [-13.67]
N	150	150	150	150
R-sq	0.916	0.295		

*t* statistics in brackets

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

(1) OLS; (2) FEM; (3) REM; (4) GLS

Sources: Authors' Calculations

The frequentist estimation results show that non-renewable energy consumption (EC) and economic growth (Gr) exert strong and positive impacts on CO<sub>2</sub> emissions in the selected ASEAN countries. Particularly, variables renewable energy (EC) and population growth (PoP) are not statistically significant in relation with CO<sub>2</sub> emissions. Contrast to the frequentist approach, where as a consequence of low statistical power, non-significant p-value cannot be meaningfully interpreted in the null hypothesis significance testing, the Bayesian estimation captures the effects of all variables (Table 3). Furthermore, in the Bayesian way, the posterior distribution is meaningful and easy to explain. The mean (regression coefficient) as a key measure of central tendency is taken to reflect a point estimate for the parameters of interest, REC, EC, Gr, and PoP. In addition, a 95% credibility interval derived from the posterior distribution implies that this interval has a 95% chance of containing the true values of the parameters, while the frequentist confidence interval only contains the true population value in 95% of the intervals over a long run of trials. Hence, the Bayesian approach provides more meaningful solutions. Table 3 demonstrates that as the credibility intervals do not contain zero, variables REC, EC, and PoP have strong positive effects on the response (CO<sub>2</sub>), while the effect of variable Gr is very weak or even ambiguous. Concerning the sign of impact, Bayesian results are similar to frequentist ones, a great difference here is that variables REC and PoP are not significant in frequentist inference.

**Table 3.** Bayesian estimation outcomes

Variable	Reg. Coeff	Std. Dev	MCSE	95% credibility interval
CO2				
REC	0.671	0.251	0.001	0.175; 1.158
EC	0.000	3.18e-06	1.8e-08	0.000; 0.000
Gr	0.050	0.045	0.000	-0.038; 0.138
PoP	0.364	0.166	0.001	0.035; 0.693
_cons	1.550	0.355	0.002	0.847; 2.247
sigma2	3.167	0.376	0.002	2.518; 3.977

Source: the authors' calculation

## 4.2 Discussions

Evidence on the role of renewable and non-renewable energy consumption in CO<sub>2</sub> emissions is consistent with studies on the same topic such as Salari et al. (2021), Bekun et al. (2019) and Yang et al. (2021). Reasons to explain the empirical results are as follows:

Most of the countries under consideration belong to the group of developing countries (except Singapore). Their economic activities depend mainly on fossil energy sources. Fossil energy usually includes oil, gas, and coal. Coal is used to generate electricity in thermal power plants and as a fuel in steam engines, locomotives; generate heat in metallurgical, cement, chemical plants. And, as we all know, fossil energy is to blame for the rise in CO<sub>2</sub> emissions in this region.

Because of the harmful effects of fossil fuels on the environment, many countries around the world, including ASEAN+5, have strengthened policies on the use of renewable energy in economic activities. This represents the contribution of both renewable and non-renewable energy in boosting GDP growth. However, the results also show that, despite recognizing the harmful effects of using fossil fuels, countries in this region still have a great dependence on them. Projects using renewable energy usually require big investment and also, machines and technique should be at high end of the field. This is seen as a barrier for developing countries.

## 5 Conclusion and Policy Implications

### 5.1 Conclusion

The study aims to estimate the impacts of renewable and non-renewable energy consumption on CO<sub>2</sub> emissions in 5 ASEAN countries for the period 1990–2019 via both frequentist and Bayesian regressions. Though both of the approaches give the similar outcomes, some non-significant variables are dropped from analysis in frequentist inference. In general, the results of the study show some notable points as follows:

Firstly, data analysis based on frequentist and Bayesian inference has shown an obvious relationship between non-renewable energy use and CO<sub>2</sub> emissions.

Secondly, non-renewable energy has a strong and positive impact on GDP per capita growth in the study area.

## 5.2 Policy Implications

Governments should implement a carbon tax policy soon. Taxing carbon is seen as a measure that both increases state budget and environmental protection gauges. Although argued so much, this measure is gradually recognized and applied by many countries in order to achieve sustainable development, reduce greenhouse gas emissions and be less dependent on fossil fuel sources. Singapore is the first one in the group of 5 ASEAN countries that has applied a tax on carbon emissions. In the opinion of the authors, this policy should be widely and synchronously applied. The imposition of a carbon tax on goods and services will reduce greenhouse gas emissions. This will push production and business entities to switch to clean energy sources. The carbon tax will also encourage manufacturer to find the most effective ways to reduce carbon emissions.

Model transformation from “linear economy” to “circular economy” should be considered a priority in the new development phase. Circular economy is turning the waste output of one industry into an input resource of another industry or circulating within an enterprise itself. The circular economy partly contributes to adding value to businesses, reducing resource exploitation, reducing waste treatment costs, and minimizing environmental pollution.

## References




- Abbasi, M.A., Parveen, S., Khan, S., Kamal, M.A.: Urbanization and energy consumption effects on carbon dioxide emissions: evidence from Asian-8 countries using panel data analysis. *Environ. Sci. Pollut. Res.* **27**(15), 18029–18043 (2020). <https://doi.org/10.1007/s11356-020-08262-w>
- Adedoyin, F.F., Zakari, A.: Energy consumption, economic expansion, and CO<sub>2</sub> emission in the UK: the role of economic policy uncertainty. *Sci. Total Env.* **738**, 140014 (2020). <https://doi.org/10.1016/j.scitotenv.2020.140014>
- Bekun, F.V., Alola, A.A., Sarkodie, S.A.: Toward a sustainable environment: nexus between CO<sub>2</sub> emissions, resource rent, renewable and nonrenewable energy in 16-EU countries. *Sci. Total Environ.* **657**, 1023–1029 (2019). <https://doi.org/10.1016/j.scitotenv.2018.12.104>
- Dehghan Shabani, Z., Shahnazi, R.: Energy consumption, carbon dioxide emissions, information and communications technology, and gross domestic product in Iranian economic sectors: a panel causality analysis. *Energy* **169**, 1064–1078 (2019). <https://doi.org/10.1016/j.energy.2018.11.062>
- Grossman, G., Krueger, A.: Environmental Impacts of a North American Free Trade Agreement. <https://EconPapers.repec.org/RePEc:nbr:nberwo:3914> (1991)
- Halicioglu, F.: An econometric study of CO<sub>2</sub> emissions, energy consumption, income and foreign trade in Turkey. *Energ. Policy* **37**(3), 1156–1164 (2009). <https://EconPapers.repec.org/RePEc:eee:enepol:v:37:y:2009:i:3:p:1156-1164>
- Hersh, M.A.: The economics and politics of energy generation. In: Kopacek, P. (ed.) *Improving Stability in Developing Nations through Automation 2006*, pp. 77–82. Elsevier (2006). <https://doi.org/10.1016/B978-008045406-1/50011-2>



- IEA: Electricity Market Report – December 2020, IEA, Paris <https://www.iea.org/reports/electricity-market-report-december-2020> (2020)
- Kasman, A., Duman, Y.S.: CO2 emissions, economic growth, energy consumption, trade and urbanization in new EU member and candidate countries: a panel data analysis. *Econ. Modell.* **44**, 97–103 (2015). <https://EconPapers.repec.org/RePEc:eee:ecmode:v:44:y:2015:i:c:p:97-103>
- Kotroni, E., Kaika, D., Zervas, E.: Environmental kuznets curve in greece in the period 1960-2014. *Int. J. Energy Econ. Policy* **10**(4), 364–370 (2020). <https://doi.org/10.32479/ijeeep.9671>
- Ozturk, I., Acaravci, A.: CO2 emissions, energy consumption and economic growth in Turkey. *Renew. Sustain. Energ. Rev.* **14**(9), 3220–3225 (2010). <https://doi.org/10.1016/j.rser.2010.07.005>
- Salari, M., Javid, R.J., Noghanibehambari, H.: The nexus between CO2 emissions, energy consumption, and economic growth in the U.S. *Econ. Anal. Policy* **69**, 182–194 (2021). <https://doi.org/10.1016/j.eap.2020.12.007>
- Sun, H., Samuel, C.A., Kofi Amissah, J.C., Taghizadeh-Hesary, F., Mensah, I.A.: Non-linear nexus between CO2 emissions and economic growth: A comparison of OECD and B&R countries. *Energy* **212**, 118637 (2020). <https://doi.org/10.1016/j.energy.2020.118637>
- Umbach, F.: ASEAN’s energy transition: Risks and opportunities. <https://www.gisreportsonline.com/r/energy-southeast-asia/> (2021)
- Verbic, M., Satrović, E., Muslija, A.: Environmental Kuznets curve in Southeastern Europe: the role of urbanization and energy consumption. *Environ. Sci. Pollut. Res.* **28**(41), 57807–57817 (2021). <https://doi.org/10.1007/s11356-021-14732-6>
- Wasti, S.K.A., Zaidi, S.W.: An empirical investigation between CO2 emission, energy consumption, trade liberalization and economic growth: a case of Kuwait. *J. Building Eng.* **28**, 101104 (2020). <https://doi.org/10.1016/j.jobbe.2019.101104>
- Yang, M., Wang, E.-Z., Hou, Y.: The relationship between manufacturing growth and CO2 emissions: does renewable energy consumption matter? *Energy* **232**, 121032 (2021). <https://doi.org/10.1016/j.energy.2021.121032>



# Machine Learning Applications on Box-Office Revenue Forecasting: The Taiwanese Film Market Case Study

Shih-Hao Lu<sup>1</sup> , Hung-Jen Wang<sup>1</sup>, and Anh Tu Nguyen<sup>2</sup>  

<sup>1</sup> Department of Business Administration, National Taiwan University of Science and Technology, Taipei 106, Taiwan  
shlu@mail.ntust.edu.tw

<sup>2</sup> Department of Banking, Ho Chi Minh City University of Banking, Ho Chi Minh City 700000, Vietnam  
tuna@hub.edu.vn

**Abstract.** The Random Forest algorithm (RFA) is used to predict the approximate final box-office revenue of a movie in the Taiwanese film market. The results show that the RFA has stable capabilities to predict the final box-office revenue of a movie during its theatrical period with an 80% overall accuracy. Two other machine learning algorithms, i.e., the Support Vector Machine and the Logistic Regression algorithms, are applied for comparison with the RFA. We find that the RFA still achieves the highest overall accuracy of prediction in our experiment. Additionally, we applied an unsupervised machine learning method to distinguish each group in the box office revenue categories in the classification problem. Also, the feature importance analysis indicates that word-of-mouth plays a vital role in theatrical revenue determination. Our findings imply several crucial suggestions for film distributors.

**Keywords:** Box-office revenue · random forest algorithm · support vector machine · logistic regression · self-organizing maps · Taiwanese film market

## 1 Introduction

The movie industry is one of the most dramatically growing industries internationally (Ghiassi, Lio, & Moon, 2015; Kim, Hong, & Kang, 2015). For instance, The Motion Picture Association of America (MPAA) reported that the box-office profits for all films had reached a high record of US\$40.6 billion and a 5% increase in sales compared to 2016 (MPAA, 2017). Similarly, the Taiwanese film market also experienced significant growth with a total of 649 movies shown in cinemas, and the total box-office gross revenue totaled approximately US\$0.34 billion with 43 million movie tickets sold in the 2017 calendar year. Nevertheless, not every film posted successful revenue (De Vany & Walls, 1999) since 30% of released movies break even and only 10% of movies make box-office profit (Hennig-Thurau, Houston, & Walsh, 2007). Therefore, from the view

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 384–402, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_49](https://doi.org/10.1007/978-3-031-35763-3_49)

of producers, distributors, and exhibitors, box-office forecasting is not only a difficult and challenging task but also an extremely important issue given that the results of these predictions directly determine their decision making (Delen, Sharda, & Kumar, 2007; Ghiassi et al., 2015; Hur, Kang, & Cho, 2016; Sharda & Delen, 2006).

A decade of research has now provided useful information on movie revenue forecasting using multiple machine learning algorithms. However, most research mainly focused on earnings for Hollywood movies or domestic earnings in the US (Brewer, Kelley, & Jozefowicz, 2009; Chintagunta, Gopinath, & Venkataraman, 2010; Dellarocas, Zhang, & Awad, 2007; Litman, 1983; Neelamegham & Chintagunta, 1999; Sawhney & Eliashberg, 1996), and other studies focused on target markets in the Chinese, Korean or Chilean film industry (Kim et al., 2015; Marshall, Dockendorff, & Ibáñez, 2013; L. Zhang, Luo, & Yang, 2009). Based on our best knowledge, total movie theater revenue and the development of forecasting models for the Taiwanese film industry using machine learning are still largely unknown and have not been investigated. Thus, we propose to employ the Random Forest algorithm (RFA) (Breiman, 2001), which has been examined to be an accurate approach in data classification (Lin, Wu, Lin, Wen, & Li, 2017; Sun, Zhong, Dong, Saeeda, & Zhang, 2017; Wu, Ye, Liu, & Ng, 2012; Ye, Wu, Huang, Ng, & Li, 2013), in our study.

This study differs from previous research based on the following factors. First, our prediction model is precise in up to 80% of all cases, which is one of the highest precision rates among box-office revenue predicting approaches to our knowledge. Second, we propose to directly collect, measure and use word-of-mouth (WOM) data to pursue feature importance, which is unaddressed in previous studies, and the final results indicate that WOM plays an active role in explaining our experiment. Finally, our study is one of the first attempts to predict theatrical revenue in Taiwan, one of the international developing markets for movies.

The paper is organized as follows. Section 2 briefly reviews the literature on forecasting box-office rentals. Section 3 provides the details of forecasting by RFA, including measurement of input variables and the RFA procedure in our experiment. The two sections that follow mainly provide empirical results along with a comparison of other approaches, including Support Vector Machine and Logistic Regression algorithm. The last two sections of the paper discuss some conclusions extracted from empirical evidence along with study limitations and further research suggestions.

## 2 Related Work

Some primary algorithm approaches, such as multiple regression models (Basuroy, Desai, & Talukdar, 2006; Brewer et al., 2009; De Vany & Walls, 1999; Duan, Gu, & Whinston, 2008a, 2008b; Elberse & Eliashberg, 2003; Eliashberg & Shugan, 1997; Litman, 1983; Litman & Kohl, 1989), Bayesian models (Ainslie, Drèze, & Zufryden, 2005; K. J. Lee & Chang, 2009; Neelamegham & Chintagunta, 1999), and machine learning algorithms (Delen & Sharda, 2010; Du, Xu, & Huang, 2014; Ghiassi et al., 2015; Hur et al., 2016; Kim et al., 2015; Sharda & Delen, 2006; L. Zhang et al., 2009; W. Zhang & Skiena, 2009), have been developed as reported in the literature on box-office revenue forecasting. Each approach has its own advantages. For example, the multiple regression models evaluate the importance of the variables, but variables must comply with the

assumption of a normal or gamma distribution. Moreover, machine learning algorithms based on assessing nonlinear forecasting do not rely on these assumptions (Hur et al., 2016).

Regarding variable importance, Litman (1983) proposed the linear regression model with eight independent variables that are grouped into three main factors to determine a movies' theatrical success: creative sphere, scheduling and release pattern, and the marketing effort. The results show that production budget, distributor, time of release, Academy Award nominations and prizes, and critic reviews have positive effects on rentals of theatrical movies. Litman and Kohl (1989) added some new variables, including well-known ideas, country of origin, market forces, and advertising budget, to the three main factors in the model of (Litman, 1983) to examine the supply-side effect on adjusted rentals of theatrical movies. The results suggest that superstar power, distributor, positive reviews, summer season, and storyline drive movie revenues. Based on the three stages of the hierarchical Bayes model, Neelamegham and Chintagunta (1999) found that several factors, such as a number of screens showing a film, local distributors' impact on cinema earnings, and genre, were similar to separate geographic area. In the general forecasting results, the mean absolute percentage error of the prelaunch model of Neelamegham and Chintagunta (1999) is 36.6% lower than the proposed model of Sawhney and Eliashberg (1996) for the U.S. market. On the other hand, approaches based on machine learning have been developed recently, and the results show the overall accuracy of forecasting. Sharda and Delen (2006) developed the artificial neural network (ANN) model to predict a movie classified into one of nine categories from Flop to Blockbuster. The results show 36.9% classification accuracy for the exact (Bingo) hit rate. They also examined the overall accuracy as determined by traditional statistical classification methods, such as Logistic Regression (30.17%), Discriminant Analysis (29.25%), and Classification and Regression Tree (31.18%). Delen and Sharda (2010) analyzed four additional classification models in additional research to enhance the results of Sharda and Delen (2006). The overall accuracy results revealed 55.49% accuracy for the Support Vector Machine, 54.62% for Random Forest, 54.05% for Boosted Tree, and 56.07% for the Fusion (average). Recently, ongoing improvements in machine learning algorithms have led to many new and fascinating applications in box-office forecasting. Ghiassi et al. (2015) employed a dynamic artificial neural network model (DAN2) and argued that DAN2 has excellent performance with 94.1% accuracy. In addition, Ghiassi et al. (2015) eliminated variables, such as competition, star value, and special effects, from DAN2 and replaced these variables with production budget, pre-released advertising expenditures, runtimes, and seasonality. As a consequence, DAN2 exhibited better performance than ANN, as assessed by Delen and Sharda (2010). W. Wang, Xiu, Yang, and Liu (2018) applied a deep belief network (DBN) model to predict box office revenue in China. The experimental results revealed that the DBN had the lowest mean absolute error and root mean square error, comparing the traditional BRP model and the back-propagation neural network. In another research for the Chinese movie market, Liao, Peng, Shi, Shi, and Yu (2020) showed that the applied stacking fusion model performed Bingo and 1-Away accuracy at 69% and 86%, respectively.

### 3 Forecasting with the Random Forest Algorithm

A two-phase study was designed to validate whether the prediction models applied in this research can evaluate variable importance robustly within the dataset and authenticate the accuracy of the proposed forecasting framework. In this study, one algorithm is employed to build the prediction model, and two additional different algorithms are utilized to compare the prediction model's performance as shown in Fig. 1.

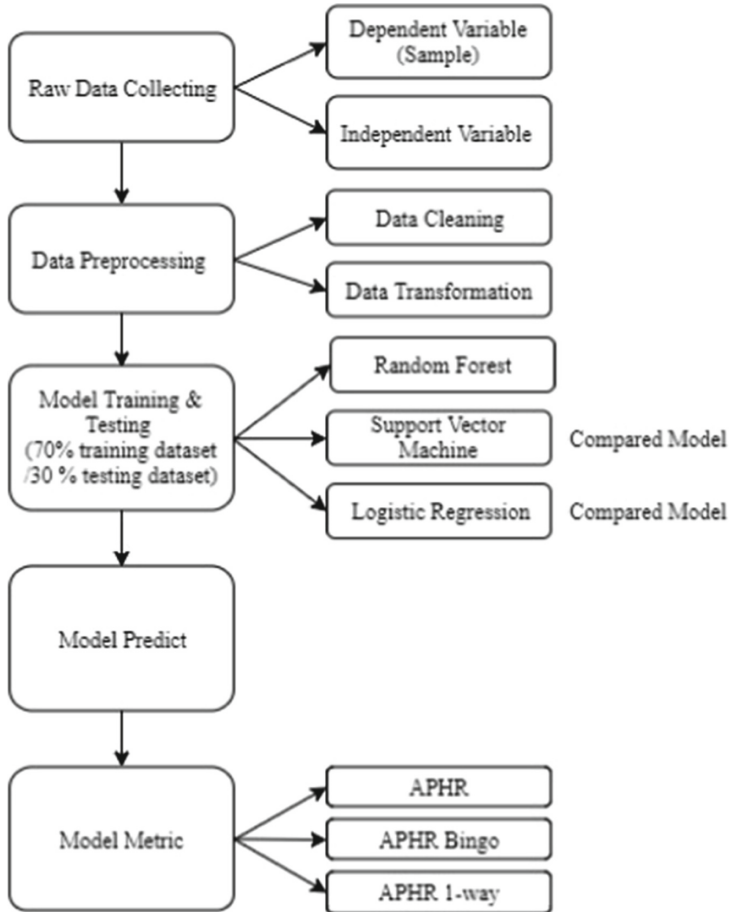


Fig. 1. Illustration of the forecasting process

#### 3.1 Raw Data Collecting

*Dependent Variable.* To validate variable importance and achieve the highest accuracy of the model, the dataset must consist of the complete calendar year information. We gathered 498 movies released between January 2017 and May 2018 from the weekly report

conducted by National Taiwan Film Institute as sample data. The dependent variable is the box-office revenue which is ranged from 3,150 New Taiwan Dollars (NT\$) to 641 million NT\$. We therefore divide box-office revenue into several categories to figure out the relationship between dependent and independent variables. However, unlike previous studies (Delen and Sharda (2010), Ghiassi et al. (2015), and Sharda and Delen (2006)) which are based on expert opinions to classify groups, we applied the self-organizing maps neural network (SOM) clustering method proposed by Kohonen (1982) to deal with group classification. SOM clustering is the unsupervised classification algorithm which is widely applied for dealing with several issues in engineering and data analysis to diagnose label of items (Markonis & Strnad, 2020; Schmidt, Rey, & Skupin, 2011). SOM clustering includes input vector and one layer of network topology which consists output neurons. The input vectors  $i = [i_1, i_2, \dots, i_n]$  are fed to the system by linear transfer function and each input node connects to the output neurons by weighted average  $w_i = [w_{i1}, w_{i2} \dots w_{in}]$ . The outcome is determined by finding a neuron which is its best matching unit (Nanda, Sahoo, & Chatterjee, 2017). The SOM has also been identified by some parameters, i.e., neighborhood area, neighborhood coefficient, and neighborhood shrinking. More specifically, the first one represents for the 2-dimensional network topology of output nodes, the second one implies for the parameter controls the interaction of output nodes inside the neighborhood area, and the third one means the neighborhood radius decreases after each iteration to figure out the best matching unit. The SOM clustering was coded by the Matlab R2019b software. The detailed procedure with all parameters set up was presented in Appendix. The results show that box-office earning in our sample should be classified into six categories as shown in Table 1.

**Table 1.** Output Variables Classified Thresholds

Class No	Range (in 10 thousands NT\$)	Numbers of samples
A	< 500 (Flop)	352
B	500–999	37
C	1,000–1,999	26
D	2,000–3,999	28
E	4,000–9,999	22
F	> 10,000 (Blockbuster)	33

*Independent Variables.* Seven different types of independent variables were used, including six variables categorized as internal variable extraction and one variable (WOM) classified as external variable extraction. With regard to the internal variables, we referred to previous studies in the literature (Ghiassi et al., 2015; Hur et al., 2016; Kim et al., 2015; Litman, 1983; Litman & Kohl, 1989; Sharda & Delen, 2006) and collect data from sources as follows.

**MPAA rating:** In the U.S., before a film is officially released on the screen, it is assigned a rating of G, PG, PG-13, R, or NC-17 based on suitability for audiences with regard to violence and sexual problems. Therefore, the input variables from The MPAA

rating system is one of the most widely utilized variables since it is an awareness system and its ratings and their definitions have remained relatively static over the past few decades (Ghiassi et al., 2015). Moreover, each particular rating decision might hold additional predictive power for box-office revenue forecasting (Ghiassi et al., 2015) because these ratings emit signals regarding film content that moviegoers find informative for personal decision making (Prag & Casavant, 1994). However, correlation results between MPAA and box-office success were divergent. Some research indicated that MPAA has a partial influence (Dellarocas et al., 2007) or even no significant influence on box-office revenue (Litman, 1983; Litman & Kohl, 1989). Meanwhile, W. Zhang and Skiena (2009) report the correlation between a movie's rating and its gross revenue. Prag and Casavant (1994) indicate that the movie ratings (G, PG, PG13, and R) have significant positive impacts on movie rental, and the MPAA ratings easily disclose movie content for films without a large budget for advertising. In this study, we use five binary variables based on Taiwan's movie rating system, which uses categories of G, P6, PG12, PG15, and R as substitutes for the MPAA ratings. Accordingly, appropriate moviegoers are classified by their ages; for example, the P6-class prohibits children under 6 and requires accompanying parents or adult guardians for children under 12.

**Genre:** Previous research commonly used movie genre as an important input variable for the theatrical success forecasting models; however, it is difficult to determine the impact of a genre on revenue because a film can be classified into multiple genres (Ghiassi et al., 2015). As a consequence, these models rarely found a significant relationship between genre and movie success or briefly classified a film based on specific genres (Litman, 1983; Litman & Kohl, 1989; Sharda & Delen, 2006; L. Zhang et al., 2009). Nevertheless, some research provided more information and pointed out that science fiction, horror, and comedy were the main movie categories associated with movie theater revenue (Litman, 1983; Liu, 2006). Furthermore, the genre of a film is an important attribute because it can determine the prospective audience demographics combined with the film rating or release timing. For example, the prospective audience for a G-rated family movie is different compared to an R-rated thriller movie, or different genres released around continuing holidays or on a particularly significant day can result in different gross revenue (Ghiassi et al., 2015). In this study, we used 19 binary-independent variables (Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Musical, Mystery, Romance, Science-Fiction, Sport, Thriller, War) that follow the convention reported by Sharda and Delen (2006) and gathered these variables from the IMDb website.

**Distributor:** To provide a more multifaceted perspective, the motion picture industry is required to make appropriate managerial decisions when distributing a film to theaters to maximize revenue (Hur et al., 2016). Moreover, the greatest distributors are more likely to announce a signal of good quality and increase film receipt (Gong, Van der Stede, & Mark Young, 2011). Litman (1983) supposed that major distributors have some advantages to produce and distribute a film to audiences given considerable financing and good connection to exhibitors' networks. As a consequence, Litman (1983) found empirical evidence that supports the hypothesis that major distributors enhance rentals of theatrical movies. In addition, Basuroy et al. (2006) reconfirmed the hypothesis of Elberse and Eliashberg (2003) that major distributors indirectly affect revenue by increasing film



screens during opening week. Generally, it can be assumed that distributor power is one of the factors that is more likely to drive a film's success. In our study, this variable is divided into three binary variables: high influence distributors, medium influence distributors, and others.

**Sequel:** In previous studies, empirical evidence showed that sequel movies correlate with the financial success of a movie (Dhar, Sun, & Weinberg, 2012; Ghiassi et al., 2015; Moon, Bergey, & Iacobucci, 2010; Sharda & Delen, 2006) even though a sequel has low quality and fewer stars (Ravid, 1999). Additionally, K. Lee, Park, Kim, and Choi (2018) noted that movie producers often produce sequel movies to reduce risk and uncertainty. In our study, the empirical model included a binary variable to identify whether a film is a sequel that is also widely used in further research (Ravid, 1999; Sharda & Delen, 2006).

**Seasonality:** An appropriate schedule for a theatrical movie might be crucial because it can likely influence the financial success of a film. One reason is that more moviegoers prefer to choose to watch a movie in their leisure time and a common film would gain great financial success during an important season, such as weekends (Duan et al., 2008a), summer months (Brewer et al., 2009; Litman, 1983), spring festival (L. Zhang et al., 2009), or the Christmas holiday (Gong et al., 2011; W. Zhang & Skiena, 2009). Commonly, film studios will schedule a theatrical movie to maximize the box-office revenue during long holidays for celebration. In this research, seasonality is measured by a binary factor that is coded 1 if a film is released on a long holiday (3 days or more) or 0 if not.

**Nationality:** Some earlier empirical studies indicated that a film origin could impact its movie theater success. For example, F. Wang, Zhang, Li, and Zhu (2010) discovered that movies produced in China have a significantly positive effect on aggregative box-office revenue. In contrast, L. Zhang et al. (2009) found that international movies can be more profitable than Chinese movies that are screened, produced, and filmed in China, and W. Zhang and Skiena (2009) showed that movies originating from the USA exhibit a clearly significant correlation with movie revenue. In Taiwan, the top ten profitable movies in 2017 were all from Hollywood. These facts suggest that a movie's financial success is more likely correlated with the place where the movie originates, and this correlation is especially significant for Hollywood movies in the Taiwanese market. In our study, we divided movie origin into seven binary variables: Hollywood, Chinese, Japanese, Korean, Thai, Bollywood, and others.

In addition to input variable features extracted from movies, the inclusion of a set of word-of-mouth (WOM) external variables in the forecasting model is imperative because it can positively influence the accuracy of the model (Asur & Huberman, 2010; Duan et al., 2008a; Liu, 2006). Liu (2006) supposed that WOM determines movie revenue in two phases. First, WOM volume increases filmgoers' awareness. Second, WOM valence affects consumers' attitudes about the films and their decisions making. However, the results in the literature are quite divergent. Some earlier studies advocate that WOM exhibits a positive contribution to the film industry (Baek, Oh, Yang, & Ahn, 2017; Du et al., 2014; Duan et al., 2008a, 2008b; Elberse & Eliashberg, 2003), whereas other studies showed a partial influence on revenue (Basuroy, Chatterjee, & Ravid, 2003; Chintagunta et al., 2010; Duan et al., 2008b; Eliashberg & Shugan, 1997; Liu, 2006). For



example, Elberse and Eliashberg (2003) argue that WOM is a crucial predictor of revenue and screens in subsequent weeks. Liu (2006) found that the volume of WOM is the most significant effect on theatrical rentals, whereas the valence of WOM is not. Similarly, Duan et al. (2008b) demonstrated that WOM valence indirectly increases box-office revenue by generating a higher volume of WOM rather than directly influencing revenue. Furthermore, scholars have recently focused on the effects of electronic WOM in the era of the Internet revolution given the speed of WOM transmission (Duan et al., 2008b). For example, Duan et al. (2008a, 2008b) tracked the data from three different social network service platforms, including Yahoo! Movies, Variety.com, and BoxOfficeMojo.com, and summarized the daily and the cumulative number of posts for each movie. Asur and Huberman (2010) collected Twitter posts per hour that mentioned a specific movie as the input data to assess the volume level. In this study, to avoid inaccurate predictions due to the omission of WOM effects, we divided the WOM variable into two components: WOM volume with three indicators (the rating of a movie, the quantity of voters for a movie, and the total views of a movie trailer) and WOM valence with two indicators (the total number of likes/dislikes for a movie trailer). WOM data are collected from IMDb, Yahoo! Movies TW, and YouTube (Table 2).

**Table 2.** Summary of Independent Variables Extracted from the Movie Aspect

Independent Variables	No. of values	Description	Data source	Independent Variables	Classification	Description	Data source
Internal Variables				External Variables			
Movie Rating	5	G, P6, PG12, PG15, R	Taiwan BAMID	The rating of a movie	Volume	Positive Integer	IMDb
Distributor	3	High, Medium influence distributors, others	NTFI	The no. of voters for a movie	Volume	Positive Integer	IMDb
Nationality	7	Hollywood, Chinese, Japanese, Korean, Thai, Bollywood, others	NTFI	The rating of a movie	Volume	Positive Integer	Yahoo! Movies TW
The Official Release Schedule	2	Continuous holiday (3 days above), No	NTFI	The no. of voters for a movie	Volume	Positive Integer	Yahoo! Movies TW
Sequel	2	Yes, No	NTFI	The total views of a movie trailer	Volume	Positive Integer	YouTube

(continued)

**Table 2.** (continued)

Independent Variables	No. of values	Description	Data source	Independent Variables	Classification	Description	Data source
Genre	19	Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Horror, Mystery, Romance, Sci-Fi, Sport, Thriller, History, Musical, War	NTFI	The total number of likes for a movie trailer	Valence	Positive Integer	YouTube
				The total number of dislikes for a movie trailer	Valence	Positive Integer	YouTube

**3.2 Data Preprocessing**

To enhance the accuracy of the model, two major data preprocessing points were applied: data cleaning and data transformation. The former relates to filling in missing values, dropping outliers, and resolving inconsistencies in the data. The latter pertains to adjusting different dimensions and increasing the accuracy of the forecasting models. In particular, data cleaning is used to remove movies merely released to the Taiwanese market because it is difficult to link and extract corresponding voting and rating data at IMDb. In addition, data cleaning is used to add the appropriate values, which are derived from a sample with similar features within the research dataset. For example, the movie ‘*Jump! Man*’ lacks the value of the number of voters on IMDb; however, the movie also belongs to the Taiwanese documentary movie dataset.

Data transformation, on the other hand, is a process that normalizes raw data with different meanings, dimensions, units, or scales to properly format data for the forecasting model. In our study, dummy variables are numbered via one-hot encoding, and the numerical variables are normalized to format the features within the raw dataset.

**3.3 Model Training and Testing**

In this study, we used the Average Percent Hit Rate (APHR), which is the most intuitive indicator to measure the discrimination for the predictive accuracy of a classification problem. We also applied two performance measures as the prediction results of the model: the average percent hit rate of exactly classifying a movie’s success (Bingo) and

1-Away. The APHR indicator, the Bingo, and the 1-Away were introduced by Sharda and Delen (2006). More specifically, the APHR measures the ratio of correct classifications to the total number of movies in the sample. The bigger values of APHR, the better the classification performance. The Bingo is applied to precisely classify a movie into one of six categories based on revenue thresholds, whereas the 1-Away is allowed for two subsequent categories, as shown in Table 1. For example, if a movie revenue is predicted to group A (revenue is less than 5 million NT\$), but the actual revenue of the movie is B (revenue is between 5 to 10 million NT\$), the precision for the Bingo is incorrect meanwhile for 1-Away is correct. In other words, the Bingo shows the exact prediction while the 1-Away allows predicted values for a broader range that may reflect real scenarios (Delen & Sharda, 2010; Ghiassi et al., 2015; Sharda & Delen, 2006).

For sample model training and dataset testing, previous studies showed that using a single experiment or a single method was inappropriate, and the subsequent use of k-fold cross-validation is ideally appropriate (Sharda & Delen, 2006). Nevertheless, K. Lee et al. (2018) demonstrated that this approach could deteriorate if the volume of data is small. If the dataset is small, repeated random subsampling validation is more suitable than k-fold cross-validation in our samples. As such, we repeat the validation process ten times using 70% of samples as training data and the remaining 30% of samples as testing data.

## 4 Results

The confusion matrix presented in Table 3 is an example of one out of the classification results of testing data subject to ten times of iteration. The columns represent the actual classes, whereas the rows represent the predicted classes in the confusion matrix. The correct classification of the samples for that class is presented in the intersection cells of the same classes.

**Table 3.** A Confusion Matrix Example of One Result from Ten-Times Iterated Classification of Testing Data

		Actual						Avg.
		A	B	C	D	E	F	
Predicted	A	109	1	1	0	0	1	
	B	7	3	0	1	0	0	
	C	5	2	0	1	0	0	
	D	1	0	2	0	2	2	
	E	1	0	0	1	2	2	
	F	0	0	0	0	0	6	
Bingo		0.89	0.50	0.00	0.00	0.50	0.55	0.41
1-Away		0.94	1.00	0.67	0.67	1.00	0.73	0.83
Average Percent Hit Rate (APHR)								0.80

Table 3 also reveals the prediction accuracy of each class individually and overall prediction accuracy in the lower column. For instance, in this iteration of the prediction

process, 109/150 samples were accurately predicted to be class A compared with real results, while others represent the misclassifications. Thus, the highest hit rates of Bingo and 1-Away for class A are 0.89 and 0.94, respectively and are the highest exact proportions among movie types. In addition, the overall APHR is 0.80, which indicates that the prediction model is able to classify 80% of samples correctly into their classes within this experiment. Furthermore, the aggregated ten-fold iterated classification results in Fig. 2 show that the average overall accuracy of APHR is 0.80, whereas the prediction accuracy of Bingo and 1-Away is 0.50 and 0.85, respectively. The obtained results are better than those in Sharda and Delen (2006), which were 37% for APHR, 37% for Bingo, and 76% for 1-Away. Although the accuracy of the Bingo metric is relatively low (50%) in our experiments, the 1-Away metric reaches 85%, indicating a practical approach for practical application, as previously mentioned.

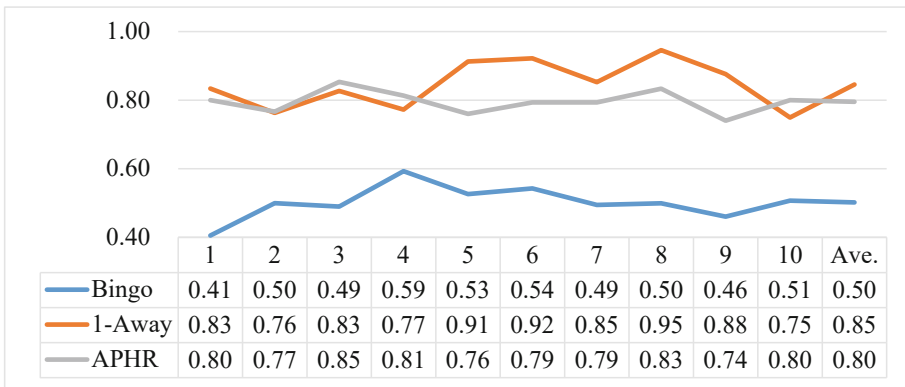


Fig. 2. An aggregated ten-times iterated classification result

To improve the accuracy of the prediction models, we perform feature importance analysis to determine the independent variable(s) that mostly affect dependent variables in the proposed models. The results are summarized in Fig. 3. Here, the x-axis represents the input variables, and the y-axis represents the importance of the input variables. Altogether, the majority of external variables (WOM) have significant contributions to the prediction of a movie’s financial success. For instance, all of these variables have important explanations for box-office revenue compared with internal variables. Additionally, most of the explanatory power is derived from the volume of WOM, which is consistent with the findings of Duan et al. (2008a, 2008b), Liu (2006). Taiwanese filmgoers mostly use information from Yahoo! Movie TW to learn about a movie.

Figure 3 also presents a point of comparison of overall APHR for RFA without internal or external variables. As the line graph suggests, the dataset without the internal variables for RFA can also result in the same overall prediction accuracy (0.8). In contrast, the dataset without the internal variables has decreased by 8%. This finding reconfirms the important contribution of WOM in our forecasting model.

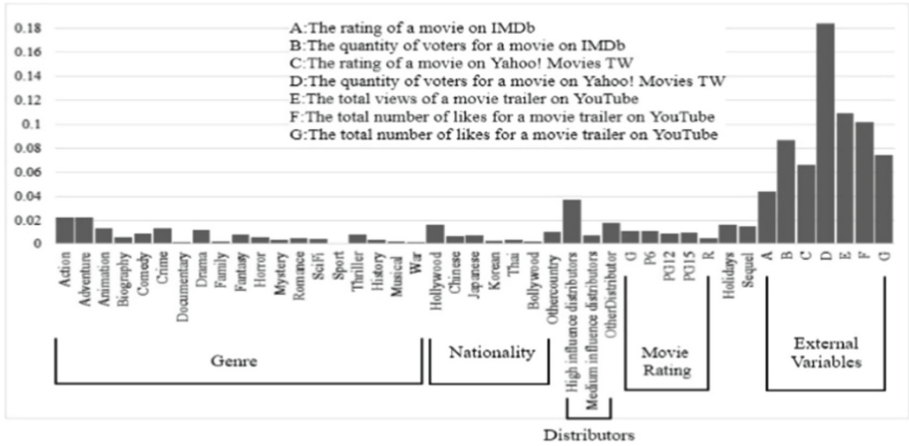


Fig. 3. Feature importance analysis

### 5 Comparison to Other Models

Two additional machine learning algorithms, Support Vector Machine algorithm (SVM) and Logistic Regression algorithm (LR), were applied to validate the performance of the prediction of the RF model. SVM is expected to identify the maximum margin hyperplanes that optimally classify the categories in the training data (Delen & Sharda, 2010; K. Lee et al., 2018), while LR is used to predict binary or multiclass dependent variables (K. Lee et al., 2018; Sharda & Delen, 2006). Using the same training and testing dataset with the same cross-validation method for RFA, SVM, and LR, Fig. 4 presents

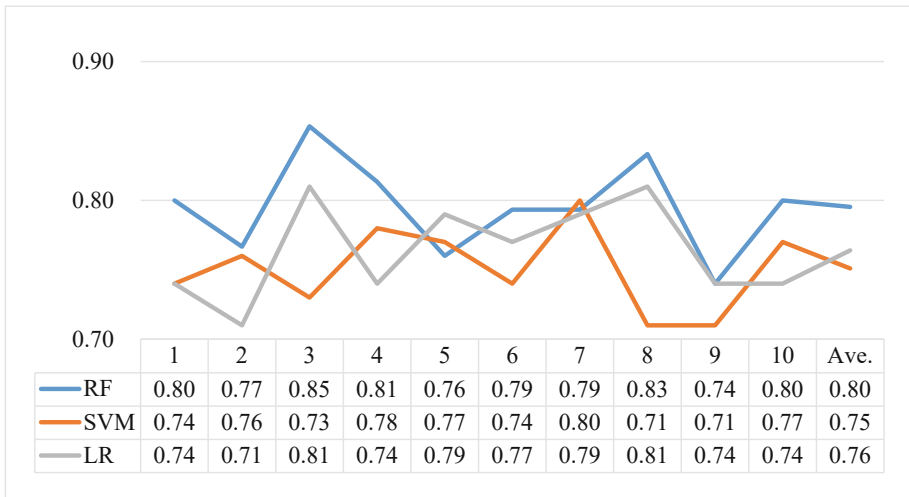


Fig. 4. An aggregated ten-times iterated classification result (APHR) for random forest, support vector machine, and logistic regression algorithm

the results for the ten-fold iteration for each approach. As the line graph shows, the RF has a higher average overall accuracy with an APHR of 0.80 than the SVM algorithm (0.75) and LR algorithm (0.76). Therefore, the RF better performs classification tasks than SVM or LR in this research. We note that the results obtained in this study are higher than those in Sharda and Delen (2006) and are the same accuracy as those in Liao et al. (2020).

## 6 Conclusion and Discussion

Some findings extracted from this study could be useful for distributors in Taiwan to determine the financial success of a movie. First, the results show that the RF has stable capabilities to predict the final box-office revenue of a movie during its theatrical period within an 80% rate of accuracy. Additionally, a comparison result of this validation demonstrates that RF achieves the highest average overall accuracy (APHR) compared to others (SVM and LR) in this research. This contribution provides a detailed framework of RF for future researchers or practical distributors in Taiwan in the field of forecasting box-office revenue. Second, in our proposed model, WOM plays a crucial role in cinema success, which is consistent with the findings of Baek et al. (2017). Given that WOM has extraordinary transmission speed through the Internet (Duan et al., 2008b) and that an appropriate marketing strategy before or after a movie's release is vital (Ghiassi et al., 2015), distributors could deploy advertising campaigns on social networks channels (IMDb, Yahoo! TV Taiwan) before a film's release to increase WOM volume or increase interaction at movie review forums (YouTube) to boost WOM valence during a film's performance at the box-office. In addition, although the internal variables have inconclusive contributions to box-office forecasting, these variables still provide different suggestions for the decision-makers to help a movie be successful. Our collected data revealed that successful movies likely have similar features. For example, if a movie is a Hollywood action or adventure movie, such as "*The Avengers*", and released on an important holiday, such as Chinese New Year or during Winter or Summer break, this movie is more likely to achieve over NT\$100 million at the box-office in the Taiwanese market.

## 7 Limitations and Future Research

We acknowledge three primary limitations involved with the use of machine learning algorithms to solve the forecasting problem within this study. The first limitation is the lack of sufficient data for the forecasting model. It is difficult to collect the exact number of total box-office sales from movie contributors, and this research relies upon data gathered from the National Taiwan Film Institute. Thus, data are potentially not accurate, and the proposed model might not reflect the real world. In addition, the movie market in Taiwan is narrow and classes B to E only represent 30% of samples compared to the A class, which represents 70% of samples. Consequently, unbalanced data might result in a reduction in the model's performance and might not produce reasonable results for these classes within this experiment. Second, concerning WOM variable reliability, the rating and the volume of voting specifically represent the audience's perspective. It is difficult to measure whether the ratings are real or fake. The results therefore could be biased. This issue is also for any forecasting model in future research. Finally, the official release schedule in Taiwan for some movies is occasionally ahead of the standard schedule in Hollywood, leading to insufficient information related to movies that can be used for the prediction models, such as the IMDb rating. Although the missing values can be completed using a variety of data preprocessing techniques, the results could be inaccurately reflected.

With regard to future research, some work is needed to improve our results. From the perspective of the variables used in the model, although our results reach 80% accuracy for predicting real cases, researchers could add different features based on the movie's aspects, i.e., the star value, the number of screens, or the special effects, to improve the final prediction results. In addition, other research suggests that the researcher can include a variety of WOM variables for comparison, such as the data compiled from Google trends (Panaligan & Chen, 2013) or other popular social media platforms. From the utilized approaches based on machine learning algorithms, we suggest that future research can explore different techniques to address the prediction problem for the movie domain within the Taiwanese market, such as a backpropagation algorithm.

## Appendix

### Self-Organizing Maps Algorithm Pseudocode for Box Office Revenue in Taiwan

---

```

1  Input Training: network parameters
2  Network parameters setting
3  initialize: 2-dimensional topology, connecting weight matrix randomly
4  iteration = 0
5  input node = 43
6  number of sample = N
7  neighborhood coefficient ( $R = 10\sqrt{2}, 8\sqrt{2}$ )
8  learning rate ( $n = 0.01, 0.1, 0.9$ )
9  for iteration: 1:100
10     for N = 1:N
11         generate zeros 2-dimensional topology
12         for j = 1: j (x-axis topology)
13             for k = 1: k (y-axis topology)
14                 for i = 1: i (input node)
15                     Calculating the net value between input and output topology
16                     Selecting the minimum node as best matching unit
17                 end
18             end
19         end
20         Calculating the output vector Y for each output layer
21         for j = 1: j (x-axis topology)
22             for k = 1: k (y-axis topology)
23                 for i = 1: i (input node)
24                      $\Delta w = n * (x(N,i) - w(i,j,k)) * \exp(-\sqrt{(j-j_{bmu})^2 + (k-k_{bmu})^2} / R)$ 
25                      $w(i,j,k) = w(i,j,k) + \Delta w(i,j,k)$ 
26                 end
27             end
28         end
29     end
30     for j = 1: j (x-axis topology)
31         for k = 1: k (y-axis topology)
32             for i = 1: i (input node)
33                 Calculating error for each input node
34                 Minimizing the sum of error
35             end
36         end

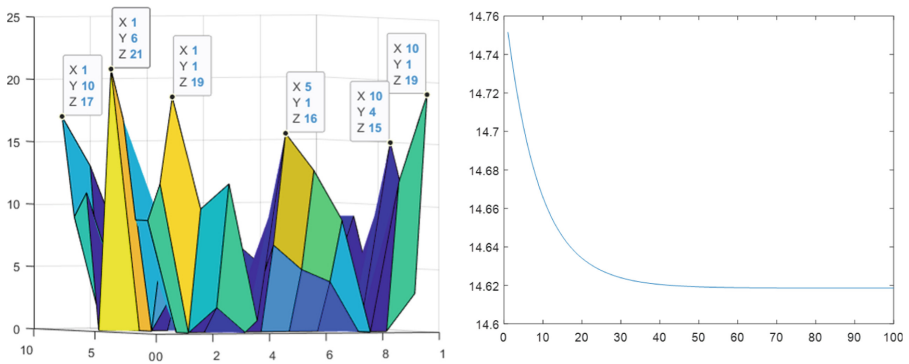
```



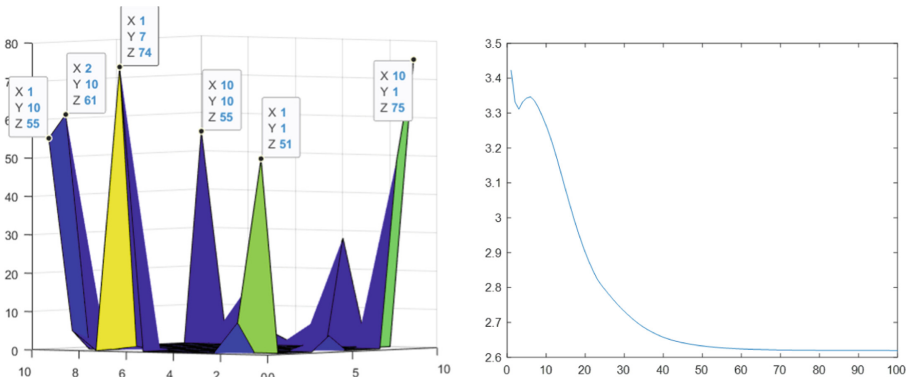
- 37        **end**  
 38        Shrinking learning rate  
 39        Shrinking neighborhood radius  
 40        **end**  
 41        **Input Recalling**  
 42        Network parameters setting as training phase  
 43        The same procedure as training phase
- 

## Self-Organizing Maps Clustering Results and Illustration of Training Error

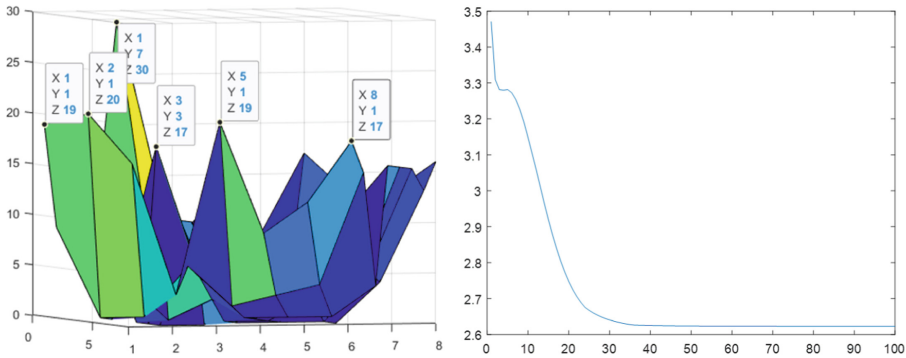
Case 1: Topology:  $10 \times 10$ , neighborhood radius:  $10 \sqrt{2}$ , shrink learning rate: 0.9; iteration: 100.



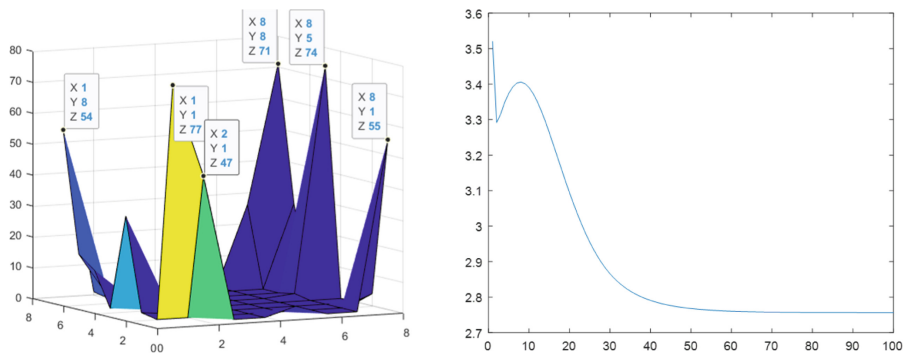
Case 2: Topology:  $10 \times 10$ , neighborhood radius:  $10 \sqrt{2}$ , shrink learning rate: 0.01; iteration: 100.



Case 3: Topology:  $8*8$ , neighborhood radius:  $8\sqrt{2}$ , shrink learning rate: 0.9; iteration: 100.



Case 4: Topology:  $8*8$ , neighborhood radius:  $8\sqrt{2}$ , shrink learning rate: 0.01; iteration: 100.



## References

- Ainslie, A., Drèze, X., Zufryden, F.: Modeling movie life cycles and market share. *Mark. Sci.* **24**(3), 508–517 (2005)
- Asur, S., Huberman, B.A.: Predicting the future with social media. In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (2010)
- Baek, H., Oh, S., Yang, H.-D., Ahn, J.: Electronic word-of-mouth, box office revenue and social media. *Electron. Commer. Res. Appl.* **22**, 13–23 (2017)
- Basuroy, S., Chatterjee, S., Ravid, S.A.: How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J. Mark.* **67**(4), 103–117 (2003)
- Basuroy, S., Desai, K.K., Talukdar, D.: An empirical investigation of signaling in the motion picture industry. *J. Mark. Res.* **43**(2), 287–295 (2006)

- Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
- Brewer, S.M., Kelley, J.M., Jozefowicz, J.J.: A blueprint for success in the US film industry. *Appl. Econ.* **41**(5), 589–606 (2009)
- Chintagunta, P.K., Gopinath, S., Venkataraman, S.: The effects of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets. *Mark. Sci.* **29**(5), 944–957 (2010)
- De Vany, A., Walls, W.D.: Uncertainty in the movie industry: does star power reduce the terror of the box office? *J. Cult. Econ.* **23**(4), 285–318 (1999)
- Delen, D., Sharda, R.: Predicting the financial success of Hollywood movies using an information fusion approach. *Indus. Eng. J.* **21**(1), 30–37 (2010)
- Delen, D., Sharda, R., Kumar, P.: Movie forecast Guru: a web-based DSS for Hollywood managers. *Decis. Support Syst.* **43**(4), 1151–1170 (2007)
- Dellarocas, C., Zhang, X.M., Awad, N.F.: Exploring the value of online product reviews in forecasting sales: the case of motion pictures. *J. Interact. Mark.* **21**(4), 23–45 (2007)
- Dhar, T., Sun, G., Weinberg, C.B.: The long-term box office performance of sequel movies. *Mark. Lett.* **23**(1), 13–29 (2012)
- Du, J., Xu, H., Huang, X.: Box office prediction based on microblog. *Expert Syst. Appl.* **41**(4), 1680–1689 (2014)
- Duan, W., Gu, B., Whinston, A.B.: Do online reviews matter?—an empirical investigation of panel data. *Decis. Support Syst.* **45**(4), 1007–1016 (2008)
- Duan, W., Gu, B., Whinston, A.B.: The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *J. Retail.* **84**(2), 233–242 (2008)
- Elberse, A., Eliashberg, J.: Demand and supply dynamics for sequentially released products in international markets: the case of motion pictures. *Mark. Sci.* **22**(3), 329–354 (2003)
- Eliashberg, J., Shugan, S.M.: Film critics: influencers or predictors? *J. Mark.* **61**(2), 68–78 (1997)
- Ghiassi, M., Lio, D., Moon, B.: Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Syst. Appl.* **42**(6), 3176–3193 (2015)
- Gong, J.J., Van der Stede, W.A., Mark Young, S.: Real options in the motion picture industry: evidence from film marketing and sequels. *Contemp. Account. Res.* **28**(5), 1438–1466 (2011)
- Hennig-Thurau, T., Houston, M.B., Walsh, G.: Determinants of motion picture box office and profitability: an interrelationship approach. *RMS* **1**(1), 65–92 (2007)
- Hur, M., Kang, P., Cho, S.: Box-office forecasting based on sentiments of movie reviews and Independent subspace method. *Inf. Sci.* **372**, 608–624 (2016)
- Kim, T., Hong, J., Kang, P.: Box office forecasting using machine learning algorithms based on SNS data. *Int. J. Forecast.* **31**(2), 364–390 (2015)
- Kohonen, T.: Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**(1), 59–69 (1982)
- Lee, K., Park, J., Kim, I., Choi, Y.: Predicting movie success with machine learning techniques: ways to improve accuracy. *Inf. Syst. Front.* **20**(3), 577–588 (2018)
- Lee, K.J., Chang, W.: Bayesian belief network for box-office performance: a case study on Korean movies. *Expert Syst. Appl.* **36**(1), 280–291 (2009)
- Liao, Y., Peng, Y., Shi, S., Shi, V., Yu, X.: Early box office prediction in China’s film market based on a stacking fusion model. *Ann. Oper. Res.* **308**, 1–18 (2020)
- Lin, W., Wu, Z., Lin, L., Wen, A., Li, J.: An ensemble random forest algorithm for insurance big data analysis. *IEEE Access* **5**, 16568–16575 (2017)
- Litman, B.R.: Predicting success of theatrical movies: an empirical study. *J. Pop. Cult.* **16**(4), 159–175 (1983)
- Litman, B.R., Kohl, L.S.: Predicting financial success of motion pictures: the ’80s experience. *J. Media Econ.* **2**(2), 35–50 (1989)
- Liu, Y.: Word of mouth for movies: its dynamics and impact on box office revenue. *J. Mark.* **70**(3), 74–89 (2006)

- Markonis, Y., Strnad, F.: Representation of European hydroclimatic patterns with self-organizing maps. *Holocene* **30**(8), 1155–1162 (2020)
- Marshall, P., Dockendorff, M., Ibáñez, S.: A forecasting system for movie attendance. *J. Bus. Res.* **66**(10), 1800–1806 (2013)
- Moon, S., Bergey, P.K., Iacobucci, D.: Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *J. Mark.* **74**(1), 108–121 (2010)
- MPAA. A comprehensive analysis and survey of the theatrical and home entertainment market environment (Theme) for 2017 - THEME REPORT (2017). [https://www.mpa.org/wp-content/uploads/2018/04/MPAA-THEME-Report-2017\\_Final.pdf](https://www.mpa.org/wp-content/uploads/2018/04/MPAA-THEME-Report-2017_Final.pdf)
- Nanda, T., Sahoo, B., Chatterjee, C.: Enhancing the applicability of Kohonen self-organizing map (KSOM) estimator for gap-filling in hydrometeorological timeseries data. *J. Hydrol.* **549**, 133–147 (2017)
- Neelamegham, R., Chintagunta, P.: A Bayesian model to forecast new product performance in domestic and international markets. *Mark. Sci.* **18**(2), 115–136 (1999)
- Panaligan, R., Chen, A.: Quantifying movie magic with google search. Google Whitepaper—Industry Perspectives+ User Insights (2013)
- Prag, J., Casavant, J.: An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *J. Cult. Econ.* **18**(3), 217–235 (1994)
- Ravid, S.A.: Information, blockbusters, and stars: a study of the film industry. *J. Bus.* **72**(4), 463–492 (1999)
- Sawhney, M.S., Eliashberg, J.: A parsimonious model for forecasting gross box-office revenues of motion pictures. *Mark. Sci.* **15**(2), 113–131 (1996)
- Schmidt, C.R., Rey, S.J., Skupin, A.: Effects of irregular topology in spherical self-organizing maps. *Int. Reg. Sci. Rev.* **34**(2), 215–229 (2011)
- Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.* **30**(2), 243–254 (2006)
- Sun, J., Zhong, G., Dong, J., Saeeda, H., Zhang, Q.: Cooperative profit random forests with application in ocean front recognition. *IEEE Access* **5**, 1398–1408 (2017)
- Wang, F., Zhang, Y., Li, X., Zhu, H.: Why do moviegoers go to the theater? The role of prerelease media publicity and online word of mouth in driving movie going behavior. *J. Interact. Advert.* **11**(1), 50–62 (2010)
- Wang, W., Xiu, J., Yang, Z., Liu, C.: A deep learning model for predicting movie box office based on deep belief network. In: Tan, Y., Shi, Y., Tang, Q. (eds.) *ICSI 2018. LNCS*, vol. 10942, pp. 530–541. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93818-9\\_51](https://doi.org/10.1007/978-3-319-93818-9_51)
- Wu, Q., Ye, Y., Liu, Y., Ng, M.K.: SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Trans. Nanobiosci.* **11**(3), 216–227 (2012)
- Ye, Y., Wu, Q., Huang, J.Z., Ng, M.K., Li, X.: Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recogn.* **46**(3), 769–787 (2013)
- Zhang, L., Luo, J., Yang, S.: Forecasting box office revenue of movies with BP neural network. *Expert Syst. Appl.* **36**(3), 6580–6587 (2009)
- Zhang, W., Skiena, S.: Improving movie gross prediction through news analysis. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (2009)



# The War on the Shadow Economy in Southeast Asia: A New Contribution from Inclusion of LGBT People

Duong Tien Ha My<sup>1</sup>, Nguyen Ngoc Thach<sup>2</sup>, Phan Thi Minh Hue<sup>3</sup>,  
and Nguyen Van Diep<sup>3</sup>✉

<sup>1</sup> Faculty of Economics and Public Management,  
Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam  
my.dth@ou.edu.vn

<sup>2</sup> Asian Journal of Economics and Banking,  
Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam  
thachnn@hub.edu.vn

<sup>3</sup> Faculty of Finance and Banking, Ho Chi Minh City Open University, Ho Chi Minh City,  
Vietnam  
{hue.ptm, diep.nv}@ou.edu.vn

**Abstract.** The inclusion of lesbian, gay, bisexual, and transgender (LGBT) individuals can have significant implications for a country's economy. This paper analyzes the effect of social inclusion of LGBT, as measured by the LGBT Global Acceptance Index (GAI), on the shadow economy size in Southeast Asia countries. To accomplish this objective, we collect data from various sources such as World Development Indicators, World Governance Indicators of the World Bank, and the Index of Economic Freedom of the Heritage Foundation over the period 2000–2017. After collecting data, we employ the Bayesian regression approach to discover the relationship between LGBT inclusion and underground activities. The result shows that an increase in the LGBT GAI will reduce the shadow economy of countries in Southeast Asia. This finding helps better understand how LGBT people thoroughly enjoy their human rights, and public attitudes towards them will contribute to the country's economic development by limiting participation in shadow economy activities. We also suggest a potential mechanism and practical implications regarding this negative relationship between LGBT inclusion and the size of the shadow economy.

**Keywords:** Shadow economy · LGBT · legal rights · Bayesian regression · discrimination

## 1 Introduction

The shadow economy is an inevitable phenomenon that profoundly impacts nations' economic, political, and social development (Lv, 2020; Nguyen and Duong, 2021; Siddik et al., 2022). Policymakers are particularly concerned about the shadow economy

because its proliferation can cause serious effects on governments, such as increased unemployment, national default risk, and at the same time decline in tax revenue and economic growth (Elgin and Uras, 2013; Nguyen and Duong, 2021; Nguyen and Duong, 2022). However, it has been very difficult to measure the shadow economy so far, and many problems of its nature and consequences have remained unexplored or fully unresolved (Siddik et al., 2022). Related studies show that economic and political factors, such as economic growth, unemployment, government size, tax burden, trade openness, and institutional quality, influence the shadow economy. Still, a large difference in the size of the shadow economy among countries remains a question in the tax compliance literature.

In recent decades, an increasing number of economists and policymakers around the world have accepted the idea that the inclusion of all groups of people in the community, especially marginalized groups, and discriminated individuals, such as LGBT people, will encourage shared prosperity and economic development (Badgett et al., 2018; Badgett et al., 2019; Vu, 2022). Specifically, Badgett et al. (2018) report a positive correlation between LGBT inclusion and GDP per capita. This finding reflects the importance of legal rights for LGBT people, and the inclusion of LGBT community explains the change in GDP per capita. Similarly, Badgett et al. (2019) also recognize that the social inclusion of LGBT people, by reflecting the legal rights and protection of the countries for LGBT individuals, positively affects average per capita income. Vu (2022) provides empirical evidence that LGBT people's social inclusion has significance for technological innovation. Particularly, Vu (2022) concludes that society's tolerance for homosexuality is positively correlated with the economic complexity index, a new measure of technological innovation at the national level. In addition, Vu (2022) also confirms that LGBT inclusion positively affects inherent human capital skills at the individual level.

In Southeast Asia, public attitudes toward LGBT are different. In a large study on "value", respondents were asked to choose types of people with whom they did not want to be neighbors. The result uncovers that the percentage of lesbian/gay people included in the list of respondents is as follows: 29.1% in Vietnam, 27.9% in the Philippines, 31.7% in Singapore, 39.8% in Thailand, 58.7% in Malaysia, and 66.1% in Indonesia. Respondents in the study said that homosexuality was never morally justified at 31.1% in the Philippines, 60.5% in Malaysia, and 87.6% in Indonesia (Manalastas et al., 2017). As for anti-LGBT criminal law, Indonesia and Malaysia have already implemented some form of criminal law for homosexual activity (Sanders, 2020). Although there have been positive changes in people's attitudes toward the LGBT community in Southeast Asia, LGBT rights in this region have been slowly improved (Sanders, 2020); this is an important issue because development agencies have focused more and more on LGBT issues, but the foundation of empirical evidence has not been still enough to direct the policy.

Based on the above point of view, this paper analyzes the influence of social acceptance for a group of LGBT people on the level of the shadow economy of nations in Southeast Asia. Social tolerance and acceptance for a group LGBT community will allow them to accumulate education and enhance health, employment opportunities, and income; these play an important role in improving the quality of human resources

in the economy. On the contrary, discrimination against a group of LGBT people can be detrimental to national human resources because the exclusion of LGBT people may force them to give up school, thereby hindering the quality of human resources and the innovation process (Kosciw et al., 2013; Vu, 2022). Furthermore, LGBT people who face discrimination in terms of employment opportunities will result in unproductive work or even unemployment (UN Human Rights, 2020). CGAP (2020) claims that LGBT people have higher unemployment rates and higher percentage of people living in poverty than the general population. Invisibility in national registries, chronic denial of reach out to health services, and high rates of violence are exacerbated by pre-existing challenges that the LGBT community faces (OAS, 2020; UNHCR, 2021). In addition, human rights experts on sexual tendency and gender identity state that the LGBT population depends on informal jobs due to challenges related to guiding documents, discrimination, and education issues. These causes motivate LGBT people to participate in shadow economy activities.

Our research contributes to the existing literature and also to practice. First, previous studies indicate that LGBT inclusion can influence the social and economic issues of a nation (Kosciw et al., 2013; Badgett et al., 2018; Badgett et al., 2019; Vu, 2022). However, very few studies examine whether and how this inclusion affects underground activities. To the best of our knowledge, the study can be the first attempt to explore the relationship between the inclusion of LGBT people and the shadow economy size using Bayesian panel data regression. Therefore, the paper would enrich research on the shadow economy. Second, we propose a mechanism through which LGBT inclusion can influence this economic sector. Third, the findings of this study help suggest some implications regarding the social inclusion of LGBT people and the underground activities management. These are two common problems in many countries.

The remainder of this paper is organized as follows: Sect. 2 presents a literature review. Section 3 describes the model, data sources, and research methods. Section 4 reports the empirical evidence. Finally, Sect. 5 presents the conclusion and policy implication.

## 2 Literature Review

### 2.1 Shadow Economy

There is currently no specific and consistent definition of the shadow economy. Even the name also shows its diversity. Many different terms are used when referring to the shadow economy, such as gray economy, ghost economy, hidden economy, invisible economy, second economy, parallel economy, informal economy, black market economy, or underground economy (Siddik et al., 2022). We will use the term “shadow economy” for this entire article. Although there are many different names, all the above names have a common meaning that is to reflect the economic sector that exists in parallel with the official economic sector.

According to the International Labor Organization (ILO), the shadow economy refers to all economic activities of workers and economic units (units in the legal sense or in practical sense) that are not regulated formally or regulated only partially (OECD & ILO, 2019). Meanwhile, Smith (1997) argues that the shadow economy is the illegal

production of goods and services that occur in the market, which is not easily detected and is not formally assessed in GDP. A similar definition proposed by OECD, ILO, IMF, and CISSTAT (2002) is that the shadow economy refers to economic activities that are not included in GDP because they are hidden and unobserved by public authorities. Those are activities that produce goods and services legally but undeclared, illegal production of goods and services and dubious income. In other words, the shadow economy may be defined as economic activities and revenues that avoid the government's regulation and tax system (Feige, 1986).

In this paper, we measure the shadow economy using the definition of Schneider et al. (2010). Accordingly, the shadow economy is market-oriented production activities that are concealed from State authorities to avoid: payment of income, added value tax, or other taxes; payment of social security contributions; must meet several legal labor market standards, minimum salary requirements, maximum working hours, safety standards, etc.; comply with certain administrative procedures, such as completing statistical questionnaires or other administrative forms (Schneider et al., 2010).

## 2.2 Theory of Discrimination

The link between LGBT inclusion and the shadow economy may be based on the theory of discrimination in labor economics. According to this theory, discriminatory employers abandon monetary profit when they refuse to hire less productive minority laborers or the minority workers that have equal productivity to the majority laborers (Becker, 1971). Without enough non-discriminatory employers, minority workers will have to do less productive work and be paid less than they are eligible for. Besides, laborers who face discrimination may be squeezed into jobs where they are less productive or may be unemployed (Bergmann, 1971). These causes will decrease the quality of human resources or inefficient use of human resources. They will be a sign of an economy that is not operating commensurate with its potential.

Many empirical studies provided the strongest evidence that LGBT people were discriminated against in job search, salary, and income (Klawitter, 2015; Neumark, 2018). Aksoy et al. (2019) showed that LGBT people were less likely to reach senior management positions, which showed the possibility of an invisible ceiling preventing the advancement of LGBT on the job ladder. In addition, discrimination against LGBT community in the educational environment also had a similar effect; specifically, discrimination would make LGBT people depressed, give up school, and have lower education levels than their capacity (Badgett et al., 2019).

According to the United Nations Development Program (UNDP), LGBT people continue to face stigma and discrimination in the Asia-Pacific region. For example, they were denied positions, denied promotions, harassed, fired or terminated, and denied partnership benefits available for opposite-sex couples (UNDP, 2019). Specifically, UNDP and ILO (2018) found that 23% of LGBT respondents in Thailand and 21% in the Philippines reported being bullied, harassed, or discriminated against in the workplace by others due to their gender identity, sexual orientation, and expression.

Badgett et al. (2021) argued that public attention to issues of equality for LGBT individuals led to major changes in non-discrimination policies in many other countries. This trend was in line with Goal 3 of the Millennium Development Goals, which



emphasized the importance of advocating for gender equality. Therefore, reducing discrimination against LGBT people was essential to achieve the third goal of the United Nations. The inclusion of the LGBT community would help them build up their human capital and allow them to reach their full potential. Those single effects are inputs to other economic processes, which implies that raising LGBT human capital and turning them more productive would generate benefits at a larger economic level.

Furthermore, creating a pro-LGBT environment around the world depends on our understanding of the economic contribution of LGBT people to social inclusion. In this paper, we argue that reducing discrimination and accepting the LGBT community will help promote economic growth by restricting LGBT people from participating in shadow economy activities. If our predictions are correct, the findings of this article at least partially support the social inclusion of LGBT community, especially in Southeast Asia, which will help reduce illegal activities and ultimately help reduce the shadow economy size of the countries in this region.

### 3 Model, Data and Methodology

#### 3.1 Model

To analyze the impact of LGBT people's social inclusion on the size of the shadow economy, we set up a model of the following form:

$$SSE_{it} = \beta_0 + \beta_1 LGBT_{it} + \sum_{k=2}^n \beta_k X_{it} + \alpha_i + \varepsilon_{it} \quad (1)$$

In which:

$SSE$  represents the size of the shadow economy.

The index  $i$  represents the country in Southeast Asia ( $i$  takes values from 1 to 9).

The index  $t$  represents the research period ( $t$  receives the period from 2000 to 2017).

$LGBT$  represents the social inclusion level of LGBT people. This variable is the main variable of interest in this article.

Parameter  $\beta_1$  records the estimated impact of LGBT adoption on the size of the shadow economy for each country.

$X$  corresponds to the set of control variables.

Parameter  $\beta_k$  is the effect of the control variables on the size of the shadow economy, respectively ( $k$  takes the value from 2 to 9).

$\alpha_i \sim N(0, \sigma_\alpha^2)$  is the individual random panel data effect.

$\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$  is the characteristic error.

#### 3.2 Data

The research sample includes nine countries in Southeast Asia (Cambodia, Indonesia, Laos, Malaysia, Myanmar, the Philippines, Singapore, Thailand, and Vietnam) during 2000–2017. These countries are selected for the study due to the availability of data.

The dependent variable is the level of the shadow economy. It is difficult to measure the level of the shadow economy, so most cross-country studies have applied the indirect

method, i.e., focusing on macroeconomic indices to estimate the size of the shadow economy. According to Lv (2020), the most common approach is the Multiple Indicators and Multiple Causes (MIMIC) model. The MIMIC approach considers the shadow economy as a latent variable, quantifying its size based on the main causes and indices of shadow activity in the economy. In this study, we use the data of Medina and Schneider (2019), who applied the MIMIC approach to estimate the size of the shadow economy of countries around the world.

We employ the Global Acceptance Index (GAI) of Flores (2019) for the LGBT-centric variable. GAI is measured based on survey data related to public beliefs and policies towards LGBT community to calculate a national acceptance score. This acceptance is the average social attitude of a country towards LGBT people as reflected in the public attitudes and beliefs about LGBT community and LGBT rights. The score of GAI ranges from 0 to 10. The lower the GAI score is, the lower the LGBT acceptance level is concerning violence and bullying, mental and physical health problems, discrimination in employment, and underrepresentation in civic leadership positions. In addition, the exclusion of LGBT people can lead to lower productivity of their employees and reduce company profits. In their empirical study, Badgett et al. (2018) used these data to measure LGBT acceptance.

In addition, in line with the result of previous studies, we will add control variables to Eq. (1).

Firstly, we use annual per capita GDP growth to measure prosperity because richer nations will have more means to monitor shadow economy activities better.

Secondly, we add the unemployment variable because unemployment often moves inversely to contraction and expansion in economic activities. Thence, when the economy suffers a significant decline and the unemployment rate rises, a number of unemployed workers can return to work in the shadow economy sector.

Thirdly, because trade openness is a proxy for the shadow economy's external dependence on the formal economy, we expect a positive correlation between trade openness and the level of the shadow economy. However, opening to international trade can also reduce the government's ability to scrutinize unofficial production. Following this argument, a negative correlation between trade openness and the size of the shadow economy may be obtained. Furthermore, trade openness can also be effective in increasing human capital. If the shadow economy is unskilled labor-intensive (and also less labor-intensive than the formal economy), this may further contribute to the shadow economy's decline (Duong et al., 2021; Elgin and Oyvatt, 2013; My et al., 2022; Siddik et al., 2022; Xu and Lv, 2022; Xu et al., 2018).

Fourthly, there is a positive correlation between tax burden and the size of the shadow economy because one of the main motives for entering the shadow sector is to avoid or evade taxes (Duong et al., 2021; My et al., 2022; Xu and Lv, 2022). However, many empirical studies (Elgin and Oyvatt, 2013; Luong et al., 2020; Torgler and Schneider, 2009) show that higher taxes will make the size of the shadow economy smaller.

Fifthly, we also control the effect of government size on the shadow economy. The linkage between government size and the shadow sector may have opposite effects. First, larger governments will have more resources to combat shadow economy activities (Siddik et al., 2022). On the other hand, a larger government size will increase excessive

government activities, which can encourage citizens and enterprises to switch to shadow economy activities (Lv, 2020; My et al., 2022; Xu and Lv, 2022; Xu et al., 2018).

Finally, we will control the impact of government effectiveness and rule of law on the shadow economy. If a government is inefficient and the level of rule of law is low, individuals and enterprises will have less trust in the government and low cooperation motive, leading them to engage in shadow activities (Luong et al., 2020; My et al., 2022; Thach et al., 2022; Torgler and Schneider, 2009). All these data are extracted from the dataset of World Development Indicators (WDI), World Governance Indicators (WGI) of the World Bank, and the Index of Economic Freedom of Heritage Foundation (2022). Table 1 presents the symbols, measurement methods, and data collection sources of the variables in the model.

**Table 1.** Variable definitions

Variable	Symbol	Measure	Source
The size of the shadow economy	SSE	Size of the shadow economy to GDP (% of GDP)	Medina & Schneider (2019)
The LGBT Global Acceptance Index	LGBT	The score of the indicator ranges from 0 to 10	Flores (2019)
GDP growth	GDPG	Annual GDP per capita growth (%/year)	World Bank (2022a)
Unemployment	UNEM	Total unemployment to total labor force (%)	World Bank (2022a)
Trade openness	OPEN	Total value of imports and exports of goods and services (% of GDP)	World Bank (2022a)
Tax burden	TAXB	The fiscal freedom index ranges from 0 (the highest tax burden) to 100 (the lowest tax burden)	Heritage Foundation (2022)
Government size	GSIZE	General government final consumption expenditure (% of GDP)	World Bank (2022a)
Government effectiveness	GEFF	The score of the indicator ranges from -2.5 to 2.5	World Bank (2022a)
Rule of law	RLAW	The score of the indicator ranges from -2.5 to 2.5	World Bank (2022a)

Table 2 summarizes two key variables, namely the level of the shadow economy and LGBT people's acceptance level in nine countries in Southeast Asia. Between 2000 and 2017, the average size of the shadow economy in Southeast Asia is 32.1% of the region's total GDP, higher than the rest of East Asia and higher than the average global rate. The level of shadow economic activities is most elevated in Cambodia and Thailand, with the

**Table 2.** Summary descriptive statistics

Country	Obs	SSE (%)	LGBT
Cambodia	18	47.39	4.60
Indonesia	18	22.98	3.03
Laos	18	28.38	5.24
Malaysia	18	29.77	4.16
Myanmar	18	46.09	4.81
Philippines	18	38.77	5.97
Singapore	18	11.18	4.99
Thailand	18	47.79	4.93
Vietnam	18	16.49	4.56
<b>Mean</b>	–	<b>32.09</b>	<b>4.70</b>

Source: The author's calculation

size of the shadow economy accounting for about 47% of GDP. In contrast, Singapore is the country with the smallest shadow economy size (accounting for about 11% of GDP).

For the LGBT variable, the LGBT people's acceptance level in nine Southeast Asian countries is higher than the average global rate (4.7 vs. 4.3). The data also show that the Philippines is the most tolerant country on LGBT issues in Southeast Asia and Asia. In contrast, Indonesia has the lowest level of LGBT acceptance in the region. This result reflects the actual growing situation of Islamic conservatism in Indonesia, leading to growing hostility towards the LGBT community (Rodríguez and Murtagh, 2022).

### 3.3 Research Method

The paper uses Bayesian regression for panel data to analyze the influence of social inclusion for the LGBT population on the size of the shadow economy. Bayesian regression is based on Bayes' theorem. This theorem can be extended to the situation of data and model parameters. With dataset  $y$  and model parameters  $\theta$ , Bayes' theorem is written as follows (van de Schoot et al., 2021):

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) \quad (2)$$

where:

$p(\theta|y)$  represents the conditional probability, where the probabilities of the model parameters ( $\theta$ ) are conditional on data ( $y$ ), also known as a posterior distribution.

$p(y|\theta)$  represents the conditional probability of the data for the model parameters, also known as the reasonable probability function.

$p(\theta)$  represents the probability that the parameter values of a particular model exist in the whole, also known as a prior distribution.

$p(y)$  is the normalization coefficient and can be discarded from Eq. (2) because it does not depend on  $\theta$ .

Therefore, the posterior distribution is proportional to the rational probability function and the prior distribution. So, Eq. (2) is simply written as  $p(\theta|y) \propto p(y|\theta)p(\theta)$  (3).

In this paper, the prior distribution used is the normal distribution and Inverse-gamma (Igamma) distribution for the parameters in the model. We assume a normal distribution for the parameters in Eq. (1) for the reasonable probability function. Finally, we use Markov Chain Monte Carlo (MCMC) method combined with the Gibbs sampling technique to generate posterior distributions. For a simulation, we run 3 Markov chains, and for each Markov chain, we take 12,500 interactions from the posterior distribution and discard the first 2,500 interactions. So the MCMC size for each chain will be 10,000.

## 4 Empirical Results

### 4.1 Bayesian Regression Result

Table 3 presents the result of Bayesian random-effects regression for panel data. Equation (1) column shows the country effects (ID) regression result. The next two columns present the country effects (ID) regression result by adding the institutional quality control variables, including rule of law and government effectiveness, respectively.

The result reveals that the LGBT variable is consistent with our expectations. The mean coefficients of LGBT variable are: -1.1636; -1.0686 and -0.9566, respectively. The probability for the LGBT variable to have a negative effect on SSE is 97%, 95%, and 94%, respectively. As such, our finding has provided strong evidence for the negative impact of LGBT community acceptance on the level of the shadow economy. Specifically, the result indicates that the higher LGBT acceptance is, the more the size of the shadow economy reduces in Southeast Asian economies. More broadly, this negative correlation is significant for economic development, meaning that the higher LGBT people's acceptance is, the better the educational level, health, and other human resources of LGBT people are. At the same time, reducing discrimination in countries with a high LGBT acceptance index will remove barriers for LGBT people, enabling them to participate fully in the formal economy and have opportunities to promote their full economic potential, thereby indirectly reducing the size of the shadow economy. This mechanism is consistent with previous studies which suggest that when LGBT individuals face challenges such as guiding documents, discrimination, and education issues, they are more inclined to join the shadow economic sector (OAS, 2020; CGAP, 2020; UNHCR, 2021).

For control variables, similar to previous studies (Luong et al., 2020; Lv, 2020; My et al., 2022; Siddik et al., 2022; Torgler and Schneider, 2009; Xu and Lv, 2022; Xu et al., 2018), we find a negative relationship between GDP growth and the size of the shadow economy. Table 3 indicates that the lowest probability of a negative effect of GDP growth on the level of the shadow economy is 86%. This result provides strong evidence that higher economic growth reduces the size of the shadow economy. Besides, our research also reports the positive effects of unemployment on the level of the shadow economy, with the probability of a positive impact on the unemployment rate being close to 100%. The finding is consistent with that of previous studies (Duong et al., 2021; Elgin and Oyvatt, 2013; Lv, 2020; Siddik et al., 2022; Torgler and Schneider, 2009; Xu et al., 2018). Another notable result is the unclear and weak effect of trade openness on the size of the shadow economy in Southeast Asian countries.

Meanwhile, the tax burden is found to have a strong positive impact on the level of the shadow economy, with the probability of a positive effect of the tax burden being close to 100%. This result is similar to the study of Duong et al. (2021), My et al. (2022), and Xu and Lv (2022). Another finding reveals that government size (expressed in government spending) has a strong negative effect on the shadow economy, with a probability of negative impact of government size being 94% or more. Higher government spending will affect salary in the formal economy, reducing the size of shadow economy activity. This result is consistent with the study of Siddik et al. (2022). Finally, we discover that government effectiveness and rule of law harm the size of the shadow economy, which is consistent with the outcome of Torgler and Schneider (2009), Luong et al. (2020), Thach et al. (2022), and My et al. (2022). However, the probability of the negative effect of rule of law on the level of the shadow economy is quite low (about 64%).

**Table 3.** Posterior simulation results

Independent variables	Equation 1	Add RLAW variable in Eq. 1	Add GEFF variable in Eq. 1
LGBT	-1.1636 (0.9682)*	-1.0686 (0.9526)*	-0.9566 (0.9366)*
GDPG	-0.1506 (0.9698)*	-0.0924 (0.8621)*	-0.1044 (0.8941)*
UNEM	1.4885 (1.0000)**	1.3753 (0.9999)**	1.2839 (0.9999)**
OPEN	0,0000 (0.5114)**	-0.0012 (0.5568)*	0.0011 (0.5489)**
TAXB	-0.2064 (0.9998)*	-0.2166 (0.9995)*	-0.1950 (0.9989)*
GSIZE	-0.2005 (0.9588)*	-0.1924 (0.9360)*	-0.1882 (0.9453)*
RLAW	–	-0.3528 (0.6365)*	–
GEFF	–	–	-1.6212 (0.8501)*
_cons	52.0044 (1.0000)**	52.9698 (1.0000)**	49.6094 (1.0000)**
ID (U0: sigma2)	298	293	259
e.SSE (sigma2)	4.1091	4.0162	3.9918

Notes: Dependent variable is SSE, in parentheses is the probability of the mean parameter's impact, \* The probability that the mean parameter has a negative effect on the SSE, \*\* The probability that the mean parameter has a positive impact on the SSE

Source: The author's calculation

## 4.2 Diagnosis of MCMC

As described in Sect. 3.3, in Bayesian estimation, the posterior distribution of the parameters is generated through the MCMC simulation method. Therefore, it is necessary to test the MCMC technique's convergence (Gelman and Rubin, 1992). In this paper, we use Gelman-Rubin's  $R_c$  statistic to diagnose MCMC convergence, a common test used for the Markov multi-chain.

Table 4 shows that  $R_c$  statistics of all parameters in the model are less than 1.1; hence there is no difference between the three Markov chains; thereby, MCMC chains have converged. This result indicates that the value of the chains is representative of the posterior distribution.

**Table 4.** MCMC diagnostics result

Independent variables	Gelman–Rubin statistic $R_c$		
	Equation 1	Add RLAW variable in Eq. 1	Add GEFF variable in Eq. 1
LGBT	1.00388	1.00779	1.01473
GDPG	1.00052	1.00059	1.00030
UNEM	1.00406	1.00739	1.01101
OPEN	1.01232	1.00200	1.00155
TAXB	1.00327	1.00452	1.00729
GSIZE	1.00108	1.00164	1.00553
RLAW	–	1.00182	–
GEFF	–	–	1.02179
_cons	1.07285	1.24118	1.05558
ID (U0: sigma2)	1.00035	1.01271	1.00934
e.SSE (sigma2)	1.00135	1.00222	1.00913

Notes: Dependent variable is *SSE*

Source: The author's calculation

## 5 Conclusion and Policy Implication

Managing the extent of the shadow economy is one of the major concerns of policymakers in countries. This paper analyzes the effect of promoting LGBT social inclusion on the level of the shadow economy of 9 nations in Southeast Asia from 2000 to 2017 period. The concern for encouraging the social inclusion of LGBT and other marginalized groups is increasing in many countries worldwide. Still, the impact of discrimination against LGBT people on economic development receives little attention from economists. The result of Bayesian regression analysis for panel data has provided evidence that LGBT people's acceptance plays an essential role in influencing the linkage between LGBT and

the shadow economy. Specifically, the result posits that the increasing LGBT community acceptance will reduce the size of the shadow economy, and this plays an indirect role in the economic development of Southeast Asian countries. In addition, we find that economic factors (GDP growth, and government size) and institutional quality (government effectiveness and the rule of law) also contribute to the reduction in the level of the shadow economy. Conversely, a high unemployment rate and tax burden increase the level of the shadow economy. Meanwhile, the effect of trade openness on the shadow economy's size is unclear.

Based on this result, we propose the following implications: Firstly, policymakers need to make efforts to improve public attitudes toward LGBT community. Secondly, policymakers should study and promulgate laws, including legal rights for LGBT. Law and acceptance can go hand in hand because they may reflect an intuitively reasonable dynamic in which the pro-LGBT public view and the introduction of pro-LGBT laws and/or rights can make the public view more supportive. Acceptance and legal rights can support each other, making the promise of legal rights into reality based on inclusion when the public view favors the LGBT community. In addition, governments of Southeast Asia nations also need policies that focus on economic factors and improve institutional quality because they will help control the size of the shadow economy.

Although the study sheds light on the relationship between LGBT inclusion and underground activities, it has some limitations. This research examines the impact of LGBT inclusion in Southeast Asian countries. However, the role of LGBT inclusion in driving the shadow economy size can vary between developing and developed countries, or across cultures. Therefore, further studies can compare the effects of this inclusion across regions to provide a more general picture. Moreover, there may be various mechanisms that link LGBT inclusion and the shadow economy. Future studies can analyze different mechanisms to better understand the relationship between these factors.

## References

- Aksoy, C.G., Carpenter, C.S., Frank, J., Huffman, M.L.: Gay glass ceilings: Sexual orientation and workplace authority in the UK. *J. Eco. Behav. Organi.* **159**, 167–180 (2019). <https://doi.org/10.1016/j.jebo.2019.01.013>
- Badgett, M.V.L., Carpenter, C.S., Sansone, D.: LGBTQ Economics. *Journal of Economic Perspectives* **35**(2), 141–170 (2021). <https://doi.org/10.1257/jep.35.2.141>
- Badgett, M.V.L., Park, E.A., Flores, A.R.: Links between economic development and new measures of LGBT inclusion. Los Angeles, CA: Williams Institute, UCLA School of Law (2018)
- Badgett, M.V.L., Waaldijk, K., Rodgers, YvdM.: The relationship between LGBT inclusion and economic development: Macro-level evidence. *World Development* **120**, 1–14 (2019). <https://doi.org/10.1016/j.worlddev.2019.03.011>
- Becker, G.S.: The economics of discrimination. University of Chicago press, Chicago, IL (1971)
- Bergmann, B.R.: The Effect on White Incomes of Discrimination in Employment. *Journal of Political Economy* **79**(2), 294–313 (1971). <https://doi.org/10.1086/259744>
- CGAP: Relief for Informal Workers: Falling through the Cracks in the COVID-19 Crisis. COVID-19 Briefing. Consultative Group to Assist the Poor. Washington, DC (2020)
- Duong, T.H.M., Nguyen, T.A.N., Nguyen, V.D.: Social capital and the shadow economy: a Bayesian analysis of the BRICS. *Asian Journal of Economics and Banking* **5**(3), 272–283 (2021). <https://doi.org/10.1108/AJEB-05-2021-0061>



- Elgin, C., Oyvatt, C.: Lurking in the cities: Urbanization and the informal economy. *Structural Change and Economic Dynamics* **27**, 36–47 (2013). <https://doi.org/10.1016/j.strueco.2013.06.003>
- Elgin, C., Uras, B.R.: Public debt, sovereign default risk and shadow economy. *Journal of Financial Stability* **9**(4), 628–640 (2013). <https://doi.org/10.1016/j.jfs.2012.09.002>
- Feige, E.L.: A Re-Examination of the “Underground Economy” in the United States: A Comment on Tanzi. *Staff Papers (International Monetary Fund)* **33**(4), 768–781 (1986). <https://doi.org/10.2307/3867216>
- Flores, A.R.: *Social Acceptance of LGBT People in 174 Countries: 1981 to 2017*. The Williams Institute, UCLA School of Law. Los Angeles, CA (2019)
- Gelman, A., Rubin, D.B.: Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**(4), 457–472 and 416 (1992). <https://doi.org/10.1214/ss/1177011136>
- Heritage Foundation: *The Heritage Foundation*. The Foundation, Washington, D.C. (2022). <https://www.heritage.org/index/>
- Klawitter, M.: Meta-analysis of the effects of sexual orientation on earnings. *Indus. Relat. A J. Eco. Soc.* **54**(1), 4–32 (2015). <https://doi.org/10.1111/irel.12075>
- Kosciw, J.G., Palmer, N.A., Kull, R.M., Greytak, E.A.: The effect of negative school climate on academic outcomes for LGBT youth and the role of in-school supports. *J. Sch. Violen.* **12**(1), 45–63 (2013). <https://doi.org/10.1080/15388220.2012.732546>
- Luong, T.T.H., Nguyen, T.M., Nguyen, T.A.N.: Rule of law, economic growth and shadow economy in transition countries. *The Journal of Asian Finance, Economics, and Business* **7**(4), 145–154 (2020). <https://doi.org/10.13106/jafeb.2020.vol7.no4.145>
- Lv, Z.: Does tourism affect the informal sector? *Annals of Tourism Research* **80**, 102816 (2020). <https://doi.org/10.1016/j.annals.2019.102816>
- Manalastas, E.J., Ojanen, T.T., Torre, B.A., Ratanashevorn, R., Hong, B.C.C., Kumaresan, V., Veeramuthu, V.: Homonegativity in Southeast Asia: Attitudes Toward Lesbians and Gay Men in Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam. *Asia-Pacific Social Sciences Review* **17**(1), 25–33 (2017)
- Medina, L., Schneider, F.: *Shedding Light on the Shadow Economy: A Global Database and the Interaction with the Official One*. Center for Economic Studies and ifo Institute (CESifo). Munich, Germany (2019)
- My, D.T.H., Vi, L.C., Thach, N.N., Van Diep, N.: A Bayesian Analysis of Tourism on Shadow Economy in ASEAN Countries. In: Ngoc Thach, N., Kreinovich, V., Ha, D.T., Trung, N.D. (eds.) *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty, and Related Topics*, pp. 405–424. Springer International Publishing, Cham (2022)
- Neumark, D.: Experimental Research on Labor Market Discrimination. *Journal of Economic Literature* **56**(3), 799–866 (2018). <https://doi.org/10.1257/jel.20161309>
- Nguyen, D.V., Duong, M.T.H.: Shadow Economy, Corruption and Economic Growth: An Analysis of BRICS Countries. *The Journal of Asian Finance, Economics and Business* **8**(4), 665–672 (2021). <https://doi.org/10.13106/jafeb.2021.vol8.no4.0665>
- Nguyen, V.D., Duong, T.H.M.: Corruption, Shadow Economy, FDI, and Tax Revenue in BRICS: A Bayesian Approach. *Montenegrin Journal of Economics* **18**(2), 85–94 (2022). <https://doi.org/10.14254/1800-5845/2022.18-2.8>
- OAS: *COVID-19: The suffering and Resilience of LGBT Persons Must Be Visible and Inform The Actions of States*. Organization of American States. Washington, D.C. (2020)
- OECD, ILO: *Definitions of informal economy, informal sector and informal employment*. OECD Publishing, Paris (2019)
- OECD, ILO, IMF, CISSTAT: *Measuring the Non-Observed Economy: A Handbook*. OECD Publishing, Paris (2002)
- Rodríguez, D.G., Murtagh, B.: Situating anti-LGBT moral panics in Indonesia. *Indonesia and the Malay World* **50**(146), 1–9 (2022). <https://doi.org/10.1080/13639811.2022.2038871>

- Sanders, D.: Sex and Gender Diversity in Southeast Asia. *Journal of Southeast Asian Human Rights* **4**(2), 357–405 (2020). <https://doi.org/10.19184/jseahr.v4i2.17281>
- Schneider, F., Buehn, A., Montenegro, C.E.: New Estimates for the Shadow Economies all over the World. *International Economic Journal* **24**(4), 443–461 (2010). <https://doi.org/10.1080/10168737.2010.525974>
- Siddik, M.N.A., Kabiraj, S., Hosen, M.E., Miah, M.F.: Impacts of Political Stability on Shadow Economy: Evidence from Bay of Bengal Initiative for Multi-sectoral Technical and Economic Cooperation Countries. *Vision* **26**(1), 221–231 (2022). <https://doi.org/10.1177/0972262920988387>
- Smith, P.M.: Assessing the size of the underground economy: The statistics Canada perspective. In: Lippert, O., Walker, M. (eds.) *The underground economy: Global evidence of its size and impact*, pp. 11–37. Vancouver, British Columbia, Canada: The Fraser Institute (1997)
- Thach, N.N., My, D.T.H., Thu, P.X., Van Diep, N.: Crime and the Shadow Economy: Evidence from BRICS Countries. In: Ngoc Thach, N., Kreinovich, V., Ha, D.T., Trung, N.D. (eds.) *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty, and Related Topics*, pp. 269–283. Springer International Publishing, Cham (2022)
- Torgler, B., Schneider, F.: The impact of tax morale and institutional quality on the shadow economy. *Journal of Economic Psychology* **30**(2), 228–245 (2009). <https://doi.org/10.1016/j.joep.2008.08.004>
- UNDP: The UNDP Asia Pacific Gender Equality Dispatch. United Nations Development Programme, New York (2019)
- UNDP, ILO: LGBTI People and Employment: Discrimination Based on Sexual Orientation, Gender Identity and Expression, and Sex Characteristics in China, the Philippines and Thailand. Bangkok: United Nations Development Programme (2018)
- UNHCR: Need to Know Guidance: Working with Lesbian, Gay, Bisexual, Transgender, Intersex and Queer Persons in Forced Displacement. UN High Commissioner for Refugees (UNHCR). Geneva, Switzerland (2021)
- UN Human Rights: COVID-19 and the Human Rights of LGBTI People. Office of the High Commissioner for Human Rights. Geneva (2020)
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Yau, C.: Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1**(1), 1–26 (2021). <https://doi.org/10.1038/s43586-020-00001-2>
- Vu, T.V.: Linking LGBT inclusion and national innovative capacity. *Social Indicators Research* **159**, 191–214 (2022). <https://doi.org/10.1007/s11205-021-02743-2>
- World Bank: World Development Indicators (WDI). The World Bank, Washington, D.C. (2022a). <https://databank.worldbank.org/source/world-development-indicators>
- World Bank: Worldwide Governance Indicators (WGI). The World Bank, Washington, D.C. (2022b). <https://info.worldbank.org/governance/wgi/>
- Xu, T., Lv, Z.: Does too much tourism development really increase the size of the informal economy? *Current Issues in Tourism* **25**(6), 844–849 (2022). <https://doi.org/10.1080/13683500.2021.1888898>
- Xu, T., Lv, Z., Xie, L.: Does country risk promote the informal economy? a cross-national panel data estimation. *Global Economic Review* **47**(3), 289–310 (2018). <https://doi.org/10.1080/1226508X.2018.1450641>



# Bayesian Hierarchical Mix-Effects Approach to Impacts of Air Pollution and Economic Growth on Private Health Care Expenditure

Bui Hoang Ngoc<sup>1</sup> and Nguyen Ngoc Thach<sup>2</sup>(✉)

<sup>1</sup> The FEMRG Research Group, Ho Chi Minh City Open University, 97 Vo Van Tan Street, District 3, Ho Chi Minh City 70000, Vietnam  
ngoc.bh@ou.edu.vn

<sup>2</sup> Banking University HCMC, Ho Chi Minh City, Vietnam  
thachnn@buh.edu.vn

**Abstract.** Assessment of the interactive effects of specific macro factors such as air pollution, economic growth, and urbanization on private health care expenditure is vital for social sustainability policies. Nevertheless, to the best of our knowledge, previous studies on this topic focused on economies outside ASEAN and, particularly, were conducted in the out-of-date frequentist framework. Furthermore, empirical outcomes are often different, even contradictory. This research aims to revisit the effects of air pollution, economic growth, and urbanization on private health care expenditure in ASEAN using the Bayesian two-level mixed-effects regression via the Metropolis-Hastings algorithm to capture varying effects of the determinants on the response variable, decreasing model uncertainty. The Bayesian outcomes show that income per capita and air pollution positively and strongly impact private health care expenditure. But, interestingly, the impact of urbanization is ambiguous.

## 1 Introduction

Over the past decade, humanity has witnessed an increase in mortality and morbidity caused by environmental pollution. Air pollution adversely affects the development of the natural world, threatening the safety of human life (Currie et al. 2008; Hansen & Selte 2000; Yazdi & Khanalizadeh 2017). The pressure of improving income per capita boosts economic activities, which intensifies the destruction of environmental quality (Hassan et al. 2019; Moutinho et al. 2017; Xu 2018). Environmental pollution and climate change trigger health problems, resulting in huge health costs. The impact of air pollution on individual health is obvious, but who pays health costs is an essential policy issue. In most emerging countries, the public health care service systems are overloaded, whereas government budgets spent on health insurance are not sufficient to meet the requirements of residents.

The nexus of economic growth, air pollution and health expenditure has been discussed in the United States and the United Kingdoms. In the early 1950s, when the American Institute of Public Health published a report of the impact of air pollution on health in 1957. Nowadays, a great number of scientists have found a positive impact of air pollution and economic growth on health spending. The pioneering work of Newhouse (1977) investigating the influence of economic growth on health care expenditure in 13 Organisation for Economic Co-operation and Development (OECD) countries revealed a positive effect of income on personal health budget. Consistent with this view, Murthy and Okunade (2016) examined the effects of the macroeconomic factors on American health spending from 1960 to 2012 by applying the ARDL approach. Real income, growth rate of the population over 65 years, and high medical technology are shown to be positively correlated to health spending. Analyzing a sample of 15 OECD countries for 1995–2011, Doğan et al. (2014) revealed that the largest and smallest effects on health care expenditure are exerted by public health expenses and labor force respectively, while the influence of income is positive. However, Hansen and King (1996) found no relationship between economic growth and health care expenses for 20 OECD countries. This conclusion was reconfirmed by McCoskey and Selden (1998). Via the fixed effect and random effect models on the data of 29 OECD countries for 1995–2014, Fernandez et al. (2019) also discovered that per capita income has a insignificantly positive effect on health spending. To explain the conflicting conclusions on the mentioned connection, Newhouse (1977), Blazquez-Fernandez et al. (2019), Liang and Tussing (2019) argued that the relationship depends on the “Norm” hypothesis, according to which physicians, medical facilities, and hospitals provide health care services by formal norms. That means health care services are offered more and better if people are involved in many types of health insurance, or participate at a high level of cost.

Like economic growth, CO<sub>2</sub> emissions tend to positively affect health expenditure. Chaabouni et al. (2016) analyzing a sample of 51 countries for 1995–2013 using the GMM approach showed that there is bi-directional causality between CO<sub>2</sub> emissions and economic growth and between health spending and economic growth for a global sample, but uni-directional causality running from CO<sub>2</sub> emissions to health spending, except for low-income countries. Employing the Wavelet method for the U.S, Alola and Kirikkaleli (2018) indicated that CO<sub>2</sub> emissions are related to healthcare. Okunade (2005) applying the ordinary least square (OLS) regression on a sample of 26 African countries found inequality and income per capita to raise health care expenditure. Mehrara et al. (2012) utilized the vector error correction model (VECM) to clarify the links between economic growth and health expenditure in 13 Middle East and North Africa (MENA) countries during 1995–2005. A bounds test confirms cointegration among the examined variables. Specifically, health care expenses have a positive impact on income growth, but with the rate of contribution declining. Using the PM10 emissions (PM10 is a particulate matter 10 μm or less in diameter, which can be drawn deep into the lungs or blood) as a proxy for environmental pollution,

Yazdi and Khanalizadeh (2017) reconfirmed the positive effect of air pollution and economic growth on health care expenditure in MENA countries. Likewise, using VECM for the Sub-Saharan African countries from 1990 to 2015, Zaidi and Saidi (2018) affirmed that economic growth positively affects health care expenses, while the effect of air pollution is negative. By using the autoregressive distributed lag (ARDL) approach to Turkey during 1975–2007, Yavuz et al. (2013) discovered that income has no effect on health care expenditure in the long-run. Furthermore, provincial output and environmental pollution have a positive impact on public health expenditure for China during 1997–2014 (Yu et al., 2016).

Recently, investigating 20 cities in China, Yang et al. (2019) found that PM<sub>2.5</sub> emissions are related to mortality, morbidity and tend to cause economic loss. In analysis of the effects of macroeconomic factors on health care expenses in 18 Arab world countries during 1995–2015, Barkat et al. (2019) specified the same model for three groups: high-, upper-middle- and lower-middle-income countries. The empirical results from the pooled mean group (PMG) and the common correlated effects (CCE) suggest that income is not the only driver of health expenditure in the Arabian countries in the long-run. Similarly, Raeissi et al. (2018) reported that a 1% increase in carbon dioxide leads to an increase of 3.32% and 1.16% in public and private health expenditures in Iran. However, Zaidi and Saidi (2018) revealed that a 1% increase in CO<sub>2</sub> and NO<sub>2</sub> emissions leads to a 0.066% and 0.577% decrease in health care expenses in the Sub-Saharan African countries. Analyzing the Greece case for 2008–2015 through a probit 2SLS regression, Kyriopoulos et al. (2019) revealed that household health care expenses respond to permanent income changes more strongly than the ones arising from current income shocks.

Similar to CO<sub>2</sub> emissions, urbanization could positively contribute to health-care. People tend to move to cities because of many opportunities for them to find a job with a good income. Still, the environmental pollution situation in big cities is getting worse and worse, and life costs there are also much higher than in villages. Rapidly developing urban areas have intensified CO<sub>2</sub> emissions (Hashmi et al. 2021). Although the relationship between air pollution, economic growth and health spending is not fully exhausted by our selective review, our paper affirms that conclusions drawn from the considered works rest mainly on data-driven frequentist inference as an obsolete estimator (Kalli & Griffin 2018; Kim 2002; Norets 2015). If the coefficients of independent variables are not yet significant, the suggestion of a conclusion or implication is difficult and even impossible (Hansen and King 1996, McCoskey and Selden 1998). The issue could be solved within the more balanced and more reliable Bayesian setting, but no research employed these methods to give substantial insights into the relationships between economic growth, air pollution, urbanization and private health care expenditure in the association of southeast Asian nations (ASEAN). This is the research gap which the current study wants to address. Hence, the contributions of the investigation are summarized as follows: first, to the best of our knowledge, ours is one of the first works accessing the impacts of air pollu-

tion, economic growth, and urbanization on private health care expenditure in ASEAN, and hence provides new empirical evidence; second, the majority of earlier studies on health care expenses utilized outdated frequency-based techniques, where coefficient parameters are fixed point estimates. More importantly, examining the links between all interested variables is impossible as non-significant coefficients are dropped out of analysis. We are the first to apply the Bayesian approach through the integrated Markov chain Monte-Carlo (MCMC) sampler to give probabilistic interpretations of model uncertainty and varying effects of air pollution, economic growth, and urbanization on private health care expenditure. With the accomplishments achieved, the research methodologically and empirically contributes to the health care expenditure field.

## 2 Methodology

In the context of an increasing crisis in standard frequentist statistics, due to the surprisingly rapid advancements of powerful computational tools, the Bayesian approach has been becoming a more and more commonly used methodology in behavioral and social sciences over the past 30 years (Lemoine 2019; Ngoc & Awan 2021). The research employs a Bayesian two-level mixed-effects method to capture the variability of initial health care expenditure across 10 ASEAN countries, due to which the precision of the estimates increases. Mixed-effects models or multilevel models are featured as incorporating both fixed effects and random effects. The former comparable to frequentist regression coefficients are estimated directly. On the contrary, the latter is summarized according to their estimated variances and covariances. Random effects may be either random intercepts or random coefficients, while the grouping structure of the data may consist of multiple levels of nested groups. We focus on random intercepts assuming that initial health care expenditure varies across the studied countries, while the effects of income, air pollution, and urbanization on private health care expenditure are the same.

We specify a random-effects model as follows:

$$\ln HE_{it} = \beta_0 + \beta_1 \ln GDP_{it} + \beta_2 \ln CO_{2,it} + \beta_3 UB_{it} + u_i + \varepsilon_{it}$$

where HE is private health care expenditure per capita (unit: US dollar), GDP is income per capita (unit: US dollar in constant 2010 prices), CO<sub>2</sub> is CO<sub>2</sub> emissions per capita (unit: metric ton), UB is urbanization rate (unit: percentage),  $u_i$  is random intercepts,  $\varepsilon$  is random error,  $i$  is country, and  $t$  is year. Annual data is collected from the World Bank database from 2000 to 2016.

Because our data sample size is sufficiently large, different prior specifications do not influence posterior results and in this situation noninformative priors are enough for modeling. For comparison purposes, informative priors for the model parameters are specified too. Accordingly, five posterior simulations are made. A sensitivity analysis to prior choice will be performed through a Bayes factor test and a model test. We assume to have models  $M_k$  parameterized by vectors

$\theta_k, k = 1, 2, \dots, r$ . Through the use of Bayes's theorem, we calculate the posterior model probabilities:

$$\rho(M_k | x) = \frac{\rho(x | M_k) * \rho(M_k)}{\rho(x)}$$

Since it is challenging to calculate  $\rho(x)$ , a popular practice is to compare two models, for example,  $M_k$  and  $M_l$  via posterior odds ratio:

$$POR_{kl} = \frac{\rho(M_k | x)}{\rho(M_l | x)} = \frac{\rho(x | M_k) * p(M_k)}{\rho(x | M_l) * p(M_l)}$$

In the case of all equally plausible models, the posterior odds ratio is transformed into the Bayes factor:

$$BF_{kl} = \frac{p(x | M_k)}{p(x | M_l)}$$

Information criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and the deviance information criterion (DIC) are utilized to determine the most suitable model among candidate models that fits the data best. The drawback of all these criteria, however, is that they either ignore prior distributions or suppose only noninformative prior distributions. They are therefore not appropriate for Bayesian sensitivity analysis, when comparison models have the same parameters but different priors. In accessing the Bayesian framework, Bayes factors are preferred to model-selection criteria as, contrast to BIC, AIC, and DIC, they contain all information about prior distributions. Focus on prior information is crucial for Bayesian sensitivity analysis, when research compares models with the same parameters but various priors. The Bayes factor of two models is just the ratio of their marginal likelihoods calculated on the same data. Bayes factors usage, nevertheless, is often criticized in some venues for a difficulty in calculating marginal likelihoods. Besides, being relative quantities, Bayes factors cannot be adopted to evaluate goodness of fit of a model of interest unless the base model fits the data well. This is another limitation of Bayes factors.

While a Bayes model test compares posterior probabilities of specified Bayesian models to determine which model is more likely among examined models on the same dataset. The present study compares the candidate models with the same parameters but different priors.

### 3 Bayesian Simulation Outcomes

#### Model Comparison

This subsection compares five posterior regression models, where the respective Gaussian prior distributions are  $N(0,1)$ ,  $N(0,10)$ ,  $N(0,100)$ ,  $N(0,1000)$ , and  $N(0,10000)$ .

**Table 1.** Bayes factor test

Model	Gaussian distribution	DIC	log(ML)	log(BF)
Model 1	N(0,1)	335.6943	-184.7604	*
Model 2	N(0,10)	336.1227	-189.1019	-4.3415
Model 3	N(0,100)	336.1339	-193.9300	-9.1696
Model 4	N(0,1.000)	336.1047	-198.5439	-13.7835
Model 5	N(0,10.000)	36.8694	-80.2871	104.4733

Note: \* means reference logBF

**Table 2.** Bayesian model test

Model	Gaussian distribution	log(ML)	P(M)	P(My)
Model 1	N(0,1)	-184.7604	0.2000	0.0000
Model 2	N(0,10)	-189.1019	0.2000	0.0000
Model 3	N(0,100)	-193.9300	0.2000	0.0000
Model 4	N(0,1.000)	-198.5439	0.2000	0.0000
Model 5	N(0,10.000)	-80.2871	0.2000	1.0000

The model comparison results are presented in Tables 1 and 2. In general, the smaller the DIC value, the larger the log(ML) and log(BF) estimate, the better a model fits the data (Table 1). P(My) denotes the posterior model probability. The higher P(My), the better a posterior model (Table 2). Consequently, model 5 is the best.

**MCMC Convergence Test**

In the application of a MCMC algorithm, convergence checks are needed before proceeding to inference. Once chain convergence is established, the model parameters have converged to equilibrium values. To avoid pseudo convergence, we simulate three MCMC chains and verify whether the results satisfy the convergence rule. This is because pseudo convergence takes place when the chains have seemingly converged, but indeed, they explored only a portion of the domain of a posterior distribution. As demonstrated in Table 3, the maximum Gelman-Rubin statistic Rc of 1.01 is close to 1.1, indicating MCMC convergence.

**Table 3.** Gelman-Rubin convergence diagnostic

Max Gelman-Rubin Rc = 1.0075	
Dependent variable: <i>lnHE</i>	Rc value
<i>lnCO<sub>2</sub></i>	1.0028
<i>lnGDP</i>	1.0006
<i>UB</i>	1.0020
<i>Intercept</i>	1.0075
<i>U<sub>0</sub> : sigma2</i>	1.0071
<i>sigma2</i>	1.0004
Convergence rule: Rc < 1.1	



The model summary reports rate of acceptance and algorithm efficiency as initial indicators of MCMC convergence. The acceptance rate is the number of proposals accepted in the total proposals, whereas algorithm efficiency is the mixing properties of MCMC sampling. Concerning the chosen model 5, the acceptance rate of 0.83 is larger than the minimum level of 0.1, whereas average efficiency is equivalent to 0.19, which is more than the acceptable level of 0.01.

**Bayesian Simulation Outcomes**

Table 4 exhibits the simulation summaries of model 5. The variables  $lnCO_2$ ,  $lnGDP$ , and  $UB$  are of our interest. With a probability of mean between 0.7 and 1,  $lnGDP$  and  $lnCO_2$  exert strongly positive effects on  $lnHE$ , while the 54% probability denotes that the impact of  $UB$  is ambiguous. The 95% credible intervals also point to similar results. The lower Monte-Carlo standard error (MCSE) values, the more accurate posterior mean estimates. For MCMC algorithms, MCSE estimates close to one decimal are acceptable. Moreover, standard deviations for all the parameters are small, indicating the preciseness of parameter estimates. Importantly, because the variations of private health care expenditure between the researched countries are captured in a Bayesian mixed-effects model, its variance estimates decrease in comparison with those from maximum likelihood estimation.

Compared to frequentist statistics, credible Bayesian intervals have direct and intuitive probabilistic interpretation. Thus, we can state that the coefficient for  $lnCO_2$  belongs to the interval  $[-0.139, 0.275]$  with a 95% probability. Similar interpretations can be offered for the remaining parameters of the model.

**Table 4.** Bayesian simulation outcomes

Variables	Mean	Std.Dev.	MCSE	Probability of mean > 0	Equal-tailed [95% Cred.Interval]
$lnCO_2$	0.0674	0.1054	0.0062	0.72	[-0.1392, 0.2747]
$lnGDP$	1.9764	0.1393	0.0048	1	[1.6994, 2.2441]
$UB$	0.0015	0.0102	0.0005	0.56	[-0.0182, 0.0218]
<i>Intercept</i>	-12.701	1.2346	0.1716	1*	[-15.139, -10.278]
$U_0 : sigma2$	5.8076	3.9268	0.2612	1	[1.8123, 15.615]
$sigma2$	0.0673	0.0088	0.0001	1	[0.0533, 0.0855]

Note: \* Probability of mean < 0

**Discussion**

Bayesian outcomes report the following findings. Per capita income, air pollution, and urbanization positively affect private health expenditure. However, their impact magnitude is different in terms of probability, with a strongly positive effect from income and a moderate effect from  $CO_2$  emission, but an ambiguous relationship between urbanization and private health care expenditure. Importantly, our estimation results are economically plausible. Compared with the previous studies, our outcome is in line with Yazdi and Khanalizadeh (2017), Zaidi and Saidi (2018), Barkat et al. (2019), reconfirming a positive impact of air pollution on health care expenditure. Regarding the effect of income on health care expenses, our result is consistent with Gerdtham et al. (1992) Hitiris

and Posnett (1992), Bhat and Jain (2006), Wang and Rettenmaier (2007). The reasons for the outcomes could be as follows: first, the rise of income allows to obtain access to better health care services at a high-cost level (Rao et al. 2009); second, the consciousness of people on their health positively changes resulting from income improvement.

## 4 Concluding Remarks

The study estimates the effects of income, air pollution, and urbanization on private health care expenditure employing the Bayesian hierarchical mixed-effects regression through the hybrid Metropolis-Hastings sampler on a panel data of 10 ASEAN countries over the period 2000 to 2016. The Bayesian mix-effects regression allows for variance reduction and so increases the accuracy of the estimates. According to the simulation outcomes, we claim in view of the probability that economic growth strongly and positively affects private health care expenditure, CO<sub>2</sub> emissions have a moderate positive impact, while the effect of urbanization is ambiguous.

Based on the obtained empirical results, two main policy implications are suggested: *First*, ASEAN countries should plan detailed policies to propagate on the harmful effects of environmental pollution, encouraging residents to use clean energy, enterprises to invest in high and green technologies; *Second*, medical waste is a factor in destroying environmental quality. Medical garbage should be well controlled by governments and hospitals.

The main limitation of this research is that we did not incorporate random coefficients due to the high dimensionality of multilevel models, which may lead to non-convergence for some model parameters.

**Acknowledgments.** We thank the anonymous reviewers and Editor-in-chief for helpful comments and suggestions. All errors are our.

## References



- Alola, A.A., Kirikkaleli, D.: The nexus of environmental quality with renewable consumption, immigration, and healthcare in the US: wavelet and gradual-shift causality approaches. *Environ. Sci. Pollut. Res. Int.* **26**(34), 35208–35217 (2019)
- Barkat, K., Sbia, R., Maouchi, Y.: Empirical evidence on the long and short run determinants of health expenditure in the Arab world. *Q. Rev. Econ. Financ.* **73**, 78–87 (2019)
- Bhat, R., Jain, N.: Analysis of public and private healthcare expenditures. *Econ. Pol. Wkly* **41**(1), 57–68 (2006)
- Blazquez-Fernandez, C., Cantarero-Prieto, D., Pascual-Saez, M.: On the nexus of air pollution and health expenditures: new empirical evidence. *Gac. Sanit.* **33**(4), 389–394 (2019)
- Chaabouni, S., Zghidi, N., Ben Mbarek, M.: On the causal dynamics between CO<sub>2</sub> emissions, health expenditures and economic growth. *Sustain. Urban Areas* **22**, 184–191 (2016)

- Currie, J., Neidell, M., Schmieder, J.: Air pollution and infant health: Lessons from New Jersey. National Bureau of Economic Research (Nber Working Paper, No. 14196) (2008)
- Doğan, İ, Tülüce, N.S., Doğan, A.: Dynamics of health expenditures in OECD countries: panel ARDL approach. *Theor. Econ. Lett.* **4**(8), 649–655 (2014)
- Gerdtham, U.-G., Sjøgaard, J., Andersson, F., Jönsson, B.: An econometric analysis of health care expenditure: a cross-section study of the OECD countries. *J. Health Econ.* **11**(1), 63–84 (1992)
- Hansen, King, A.: The determinants of health care expenditure: a cointegration approach. *J. Health Econ.* **15**(1), 127–137 (1996)
- Hansen, A.C., Selte, H.K.: Air pollution and sick-leaves: a case study using air pollution data from Oslo. *Environ. Resource Econ.* **16**(1), 31–50 (2000)
- Hashmi, S.H., Fan, H., Habib, Y., Riaz, A.: Non-linear relationship between urbanization paths and CO<sub>2</sub> emissions: a case of South, South-East and East Asian economies. *Urban Climate* **37**, 100814 (2021)
- Hassan, S.T., Xia, E., Khan, N.H., Shah, S.M.A.: Economic growth, natural resources, and ecological footprints: evidence from Pakistan. *Environ. Sci. Pollut. Res. Int.* **26**(3), 2929–2938 (2019)
- Hitiris, T., Posnett, J.: The determinants and effects of health expenditure in developed countries. *J. Health Econ.* **11**(2), 173–181 (1992)
- Kalli, M., Griffin, J.E.: Bayesian nonparametric vector autoregressive models. *J. Econ.* **203**(2), 267–282 (2018)
- Kim, J.-Y.: Limited information likelihood and Bayesian analysis. *J. Econ.* **107**(1–2), 175–193 (2002)
- Lemoine, N.P.: Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos* **128**(7), 912–928 (2019)
- Liang, L.L., Tussing, A.D.: The cyclicalities of government health expenditure and its effects on population health. *Health Policy* **123**(1), 96–103 (2019)
- McCoskey, S.K., Selden, T.M.: Health care expenditures and GDP: panel data unit root test results. *J. Health Econ.* **17**(3), 369–376 (1998)
- Mehrrara, M., Fazaeli, A.A., Fazaeli, A.A., Fazaeli, A.R.: The relationship between health expenditures and economic growth in Middle East & North Africa (MENA) countries. *Int. J. Bus. Econ. Res.* **3**(1), 425–428 (2012)
- Moutinho, V., Varum, C., Madaleno, M.: How economic growth affects emissions? An investigation of the environmental Kuznets curve in Portuguese and Spanish economic activity sectors. *Energy Policy* **106**, 326–344 (2017)
- Murthy, V.N.R., Okunade, A.A.: Determinants of U.S. health expenditure: evidence from autoregressive distributed lag (ARDL) approach to cointegration. *Econ. Model.* **59**, 67–73 (2016)
- Newhouse, J.P.: Medical-care expenditure: a cross-national survey. *J. Hum. Resour.* **12**(1), 115–125 (1977)
- Ngoc, B.H., Awan, A.: Does financial development reinforce ecological footprint in Singapore? Evidence from ARDL and Bayesian analysis. *Environ. Sci. Pollut. Res. Int.* **29**(6), 24219–24233 (2021)
- Norets, A.: Bayesian regression with nonparametric heteroskedasticity. *J. Econ.* **185**(2), 409–419 (2015)
- Raeissi, P., Harati-Khalilabad, T., Rezapour, A., Hashemi, S.Y., Mousavi, A., Khodabakhshzadeh, S.: Effects of air pollution on public and private health expenditures in Iran: a time series study (1972–2014). *J. Prev. Med. Public Health* **51**(3), 140–147 (2018)

- Rao, R.R., Jani, R., Sanjivee, P.: Health, quality of life and GDP: an ASEAN experience. *Asian Soc. Sci.* **4**(4), 70–76 (2009)
- Wang, Rettenmaier, A.J.: A note on cointegration of health expenditures and income. *Health Econ.* **16**(6), 559–578 (2007)
- Xu, T.: Investigating environmental Kuznets curve in China-aggregation bias and policy implications. *Energy Policy* **114**, 315–322 (2018)
- Yang, S., Fang, D., Chen, B.: Human health impact and economic effect for PM2.5 exposure in typical cities. *Appl. Energy* **249**, 316–325 (2019)
- Yazdi, K.S., Khanalizadeh, B.: Air pollution, economic growth and health care expenditure. *Econ. Res.-Ekonomiska Istraživanja* **30**(1), 1181–1190 (2017)
- Zaidi, S., Saidi, K.: Environmental pollution, health expenditure and economic growth in the Sub-Saharan Africa countries: panel ARDL approach. *Sustain. Urban Areas* **41**, 833–840 (2018)



# Market Share Forecast of Vietnam and of the World's Leading Textile and Garment Exporters by VAR Bayesian Model

Nguyen Thi Ngoc Diep<sup>1,2</sup> , Tran Quang Canh<sup>3</sup> , and Nguyen Ngoc Thach<sup>4</sup>

<sup>1</sup> University of Economics and Law, Ho Chi Minh City, Vietnam  
diepntn@uel.edu.vn

<sup>2</sup> Vietnam National University, Ho Chi Minh City, Vietnam

<sup>3</sup> Faculty of Business Administration, Ho Chi Minh City University of Economics and Finance,  
Ho Chi Minh City, Vietnam  
canhtq@uef.edu.vn

<sup>4</sup> Banking University of Ho Chi Minh City, 36 Ton That Dam Street, District 1, Ho Chi Minh  
City, Vietnam  
thachnn@buh.edu.vn

**Abstract.** This article forecasts the export market share of textiles during the 2020–2030 period in 10 countries with the world's largest export market share of textiles in 2019. Using the Var Bayesian model and secondary data collected from 1988 to 2019, the export market share of textiles and garments in the world in 2019, as reported on the World Bank's website, includes Vietnam, the USA, Germany, Japan, China, England, France, Italy, the Netherlands, and Spain. The forecast results show that the export market share of textiles in four countries, including the USA, China, Japan, and Spain, tends to increase slightly and steadily. In comparison, the second group of six countries, including Germany, England, France, Italy, the Netherlands, and Vietnam, tends to reduce. From a policy perspective, the authors also offer some recommendations for how textile export enterprises should operate. They should actively prepare for production and human resources after the wave of employees leaving industry zones to avoid the pandemic, bringing effective manufacturing and exporting activities and maintaining the export market share of textiles. As a means of minimizing the harmful effects of the COVID – 19 epidemic, policymakers are called upon to diversify export markets, actively search for new markets, take advantage of the Free Trade Agreement, and implement harmonized dual – target strategies in order to promote economic growth.

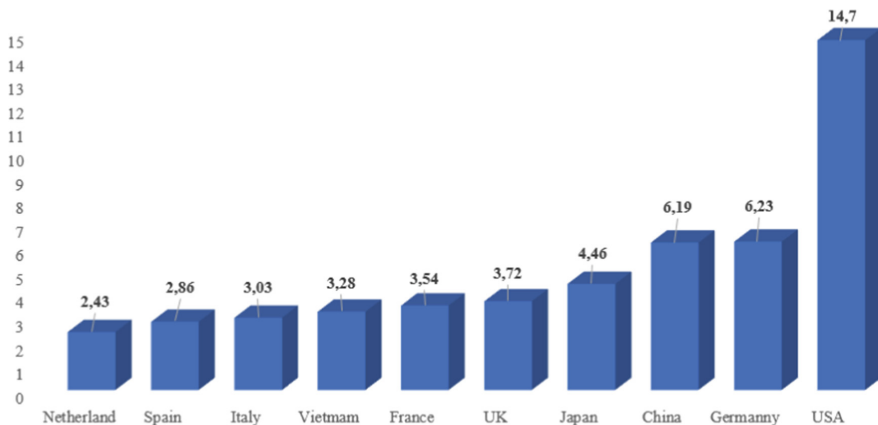
**Keywords:** Textile · VAR Bayesian model · Vietnam

## 1 Introduction

Textile industry in Vietnam has a critical position in the development process, significantly contributing to socio – economic growth. This industry is not only to serve people's daily needs but also to create jobs for tens of thousands of domestic workers,

promote exports, and increase foreign investment attraction. Moreover, the textile industry promotes the development and is an input for many other economic sectors, such as agriculture and auxiliary industries. In recent years, Vietnam's textile and garment industry has taken a positive step; the increase in the average industrial production of the textile industry index from 2012 to 2020 reached 11.8% per year. Vietnam's textile and garment exports reached 29.81 billion USD, down 9.2% compared to 2019. Like other industries, Vietnam's textile and garment industry has also been negatively impacted by the COVID – 19 pandemic, which has seriously affected production, broken the supply of raw materials, and narrowed the market for garments. When consumers were only interested in essential utensils and epidemic prevention, the demand for textile products worldwide decreased.

Along with the efforts to research and produce successful vaccines and drugs for the treatment because of COVID – 19 in developed countries such as the USA, England, Russia, and China, the governments of countries worldwide have flexible reactions to limit the spread, gradually restricting the number of deaths while opening progressively to restore the economy, creating a psychology that works for workers in all economic sectors, including Vietnam's textile and garment industry. Specifically, in June and in the first six months of 2021, the textile and garment industry has more prosperous signals than last year, thanks to the production string recovering and traditional orders rising again. Consumer shopping needs in the US and Europe for clothing increased strongly when the economy was restored due to the gradual removal of the blockade. Some major export markets of Vietnam are progressively recovering and taking advantage of opportunities from signed free trade agreements (FTAs) and going into execution.



**Fig. 1.** Top 10 countries with textile and garment export market share in the world in 2019. Source: <https://www.wto.org/>

According to the WTO 2019 (Textiles and Clothing Exports by Country & Region US\$000 2015 | WITS Data, n.d.), Vietnam's fabric export marketplace percentage is presently ranked seventh worldwide (Fig. 1). By the end of 2020, in line with the General Statistics Office of Vietnam (Ha, 2021a, 2021b), the markets of the United States, EU,

Japan, and Korea will have many high-quality symptoms, while many fabric firms will have export orders until the end of the first quarter of 2021. The Vietnam fabric enterprise index in June 2021 improved by 3.8% over the preceding month and 14.3% compared to the identical period in 2020. Moreover, the dress enterprise index improved by 3.4% and 7.9% compared to the preceding month and the identical period in 2020, respectively. In the first six months, the fabric manufacturing index improved by 8.6% compared to last year's period, and the dress enterprise improved by 8.9%. According to the Ministry of Industry and Trade (Ministry of Industry and Trade, 2021), Vietnam accounts for more than 20% of the market share of garments in the United States for the first time in decades. The reason is that garment labels have been transferring orders from China to Vietnam to avoid the influence of the US – China trade conflict.

The global textile market tends to recover due to economic support packages. With positive information on the deployment of the COVID – 19 epidemic prevention vaccine, the demand for items in general and apparel, in particular, has partly recovered. In addition, in 2021, textile enterprises were okay with a lack of raw materials. Large orders, combined with the advantages of the export market, have helped Vietnamese textiles gradually recover with export turnover growth every month while providing Vietnamese textile enterprises the flexibility to adapt to new business conditions. In particular, as the enterprise continues to invest in equipment, automation technology is also one factor contributing to the textile and garment industry's foundation to withstand the market's pressure for quality and delivery fast (Ministry of Industry and Trade, 2021).

In 2022, with the expectation of life returning to normal, the needs of people's shopping after a repressed year will grow again. This expectation helps fashion brands become more optimistic about business prospects, positively impacting traditional garment orders at Vietnamese factories. However, to improve production and business results, businesses must adjust, follow market demand, and move faster with market fluctuations to take advantage of opportunities. Businesses have been more concerned about the construction of domestic fabric supplies to exploit the advantages of the Vietnam – EU Free Trade Agreement (EVFTA), the Comprehensive Partner and Trans-Pacific Partnership Agreement (CPTPP), and the Vietnam Free Trade Agreement with England (UKVFTA). In particular, many businesses have exploited domestic production advantages to export to Europe, enjoying the advantage according to EVFTA. According to businesses, the opportunity to increase European exports has created a catalyst for manufacturers to invest in factories boldly and prepare and supply raw materials "Made in Vietnam" more (Ha, 2021a, 2021b).

This study was conducted to forecast the world's market share and the world's leading textile and garment exporters by approaching VAR Bayes to report the textile import demand. Vietnam and ten countries have the leading textile exports in the world (according to WTO data in 2019), including the US, Japan, China, Germany, England, France, Italy, the Netherlands, Spain, and Vietnam. Research results will help Vietnamese textile enterprises identify their export market share in the coming years. Research results also provide appropriate policies to reserve capital and raw materials and attract textile human resources to ensure export volume, especially if they can respond promptly and ensure production progress with the effects of the new species of SARS – CoViD – 2

virus. Part 2 of the article shows the theoretical basis, Part 3 is a research method, Part 4 is the study result, and Part 5 is the policy implications.

## 2 Literature Review

The export market for ready-made garments (RMG) is led by developing countries, mainly China, Bangladesh, Vietnam, and India. Many countries are considering the emerging economies expected to lead the world in the coming decades (Barua et al., 2018). China is predicted to overtake the US as the most prominent global fashion market. Middle – class consumers in India are part of a growing market. India is growing as producers are also becoming increasingly shrewd at technological innovation (Megersa, 2019).

Textiles are one of the traditional products that not only bring high socio – economic value to Vietnam but also contribute significantly to the process of industrialization and modernization of the country. Vietnam’s textile and garment exports in recent years have continuously increased in the number, type, and value of export turnover, making Vietnam one of the ten countries with the most prominent textile export turnover (Data in 2019).

Vietnam’s accession to the WTO in 2010 It can be said that when entering the world textile and garment market, especially the EU, Japan, and US markets by export, the biggest and most formidable competitor for Vietnamese textile and garment enterprises is China (Do, 2021).

The increased market power of Vietnamese producers depended largely on global buyers’ views of the risks of sourcing from China (Goto et al., 2011). Vietnam is a supply of 25% of goods for VF Corporation (VFC), the world’s largest garment company (according to figures in 2018), with subsidiaries including Vans, the North, Timberland, Wrangler, and Eastpak (Megersa, 2019).

## 3 Methodology

### 3.1 Data

Research data shows the export market share of textiles and garments of the ten countries with the largest market share in the world in 2019 on the Worldbank’s website. Secondary data collected from 1988 to 2019 covers the share of textile and garment exports of the US, Germany, Japan, China (including Hong Kong), the UK, France, Italy, the Netherlands, Spain, and Vietnam (wits.worldbank, n.d.)

### 3.2 Forecast Method

#### 3.2.1 Bayesian Var Model

In this article, we used the Bayesian Var model. Bayes’ analysis requires knowledge about pre – distribution attributes, capabilities, and later (Kreinovich et al, 2019; Nguyen T.N et al., 2019; Thach N.N et al., 2021; Thach N.N et al., 2022). Previously, external distribution information was based on the trust of researchers about the concerned parameters.



Possibly, the data information is available in the sample probability distribution function (PDF). Combine previous distribution through Bayes' theorem with data capabilities leading to the subsequent distribution. In particular, denote the parameters of interest in a given model by  $\theta = (\beta, \Sigma_\epsilon)$ , and the data by  $y$ . The prior distribution is  $\pi(\theta)$ , and the likelihood is  $l(y|\theta)$ , then the posterior distribution  $\pi(y|\theta)$  is the distribution of  $\theta$  given the data  $y$  and may be derived by

$$\pi(y|\theta) = \frac{\pi(\theta)l(y|\theta)}{\int \pi(\theta)l(y|\theta)d\theta}$$

To relate this general framework to Bayesian VAR (BVAR) models, suppose that we have the VAR(p) model:

$$y_t = a_0 + \sum_{j=1}^p A_j y_{t-j} + \epsilon_t \tag{1}$$

where  $y_t$  for  $t = 1, \dots, T$  is an  $m \times 1$  vector containing observations on  $m$  v different series and  $\epsilon_t$  is an  $m \times 1$  vector of errors where assume  $\epsilon_t$  is i.i.d.  $N(0, \Sigma_\epsilon)$ . For compactness we may rewrite the model as:

$$Y = XA + E \tag{2}$$

or

$$y = (I_m \otimes X)\theta + e \tag{3}$$

where  $Y$  and  $E$  are  $T \times m$  matrices and  $X = (x_1, \dots, x_T)'$  is a  $T \times (mp + 1)$  matrix for  $x_t = (1, y'_{t-1}, \dots, y'_{t-p})$ ,  $I_m$  is the identify matrix of dimension  $m$ ,  $\theta = vec(A)$ , and  $e \sim N(0, \Sigma_\epsilon \otimes I_T)$ . Using Eq. (3) the likelihood function is

$$l(\theta, \Sigma_\epsilon) \propto |\Sigma_\epsilon \otimes I_T|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - (I_m \otimes X)\theta)' (\Sigma_\epsilon \otimes I_T)^{-1} (y - (I_m \otimes X)\theta) \right\} \tag{4}$$

Let assume  $\Sigma_\epsilon$  is known and a multivariate normal prior for  $\theta$ :

$$\Pi(\theta) \propto |V_0|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\theta - \theta_0)' V_0^{-1} (\theta - \theta_0) \right\} \tag{5}$$

where  $\theta_0$  is the prior mean and  $V_0$  is the prior covariance. When we combine this prior with the likelihood function in Eq. (4), the posterior density can be written as

$$\Pi(\theta|y) = \exp \left\{ -\frac{1}{2} ((V_0^{-1/2}(\theta - \theta_0))' (V_0^{-1/2}(\theta - \theta_0)) + \left\{ (\Sigma_\epsilon^{-1/2} \otimes I_T)y - (\Sigma_\epsilon^{-1/2} \otimes X)\theta \right\}' \left\{ (\Sigma_\epsilon^{-1/2} \otimes I_T)y - (\Sigma_\epsilon^{-1/2} \otimes X)\theta \right\} \right\} \tag{6}$$

which is a multivariate normal pdf For simplicity, define

$$\omega \equiv \begin{bmatrix} V_0^{-1/2} \theta_0 \\ (\Sigma_\epsilon^{-1/2} \otimes I_T)y \end{bmatrix}$$

$$W \equiv \begin{bmatrix} V_0^{-1/2} \\ (\Sigma_\epsilon^{-1/2} \otimes X) \end{bmatrix} \tag{7}$$

then the Exponent in Eq. (6) can be written as

$$\begin{aligned} \Pi(\theta|y) \propto \exp\left\{-\frac{1}{2}(\omega - W\theta)'(\omega - W\theta)\right\} \propto \\ \exp\left\{-\frac{1}{2}(\theta - \bar{\theta})'W'W(\theta - \bar{\theta}) + (\omega - W\bar{\theta})'(\omega - W\bar{\theta})\right\} \end{aligned} \tag{8}$$

where the posterior mean  $\bar{\theta}$  is

$$\bar{\theta} = (W'W)^{-1}W'\omega = \left[V_0^{-1} + (\Sigma_\epsilon^{-1} \otimes X'X)\right]^{-1} \left[V_0^{-1}\theta_0 + (\Sigma_\epsilon^{-1} \otimes X)'y\right] \tag{9}$$

Since  $\Sigma_\epsilon$  is known, the second term of Eq. (8) has no randomness about  $\bar{\theta}$ . The posterior there fore may be summarized as

$$\begin{aligned} \pi(\theta|y) \propto \exp\left\{-\frac{1}{2}(\theta - \bar{\theta})'W'W(\theta - \bar{\theta})\right\} \\ = \exp\left\{-\frac{1}{2}(\theta - \bar{\theta})'\bar{V}^{-1}(\theta - \bar{\theta})\right\} \end{aligned} \tag{10}$$

and the posterior covariance  $\bar{V}$  is given as

$$\bar{V} = \left[V_0^{-1} + (\Sigma_\epsilon^{-1} \otimes X'X)\right]^{-1} \tag{11}$$

### 3.2.2 KPSS (Kwiatkowski-Phillips-Schmidt-Shin Test)

First, test for stationarity of the time series by KPSS test.

The KPSS test differs from the other unit root tests described here in that the series is assumed to be (trend – ) stationary under the null. The KPSS statistic is based on the residuals from the OLS regression of  $Y_t$  on the exogenous variables  $x_t$ :

$$y_t = x_t'\delta + u_t \tag{12}$$

The LM statistic is be defined as:

$$LM = \sum S \frac{(t)^2}{(t^2f_0)} \tag{13}$$

where, is an estimator of the residual spectrum at frequency zero and where is a cumulative residual function:

$$S(t) = \sum_{r=1}^t u_r \tag{14}$$

Based on the residuals  $\hat{u}_t = u_t - x_t'\delta(0)$ . We point out that the estimator of used in this calculation differs from the estimator for used by GLS detrending since it is based on a regression involving the original data and not on the quasi-differenced data.

### 3.2.3 Parameters for Bayesian VAR Model

Set parameters for Bayesian VAR model:

*The Prior type:* Litterman/Minnesota:

Early work on Bayesian VAR priors was done by researchers at the University of Minnesota and the Federal Reserve Bank of Minneapolis (see Litterman, 1986; and Doan, Litterman, and Sims, 1984), and these early priors are often referred to as the “Litterman prior” or the “Minnesota prior.” This family of priors is based on the assumption that  $\Sigma_\epsilon$  is known; replacing  $\Sigma_\epsilon$  with its estimate  $\hat{\Sigma}_\epsilon$ . This assumption yields simplifications in prior elicitation and computation of the posterior. In this paper, we choose the diagonal VAR as the estimator of  $\Sigma_\epsilon$ .

Diagonal VAR:  $\hat{\Sigma}_\epsilon$  is restricted to be a diagonal matrix (as in the univariate VAR estimator). However, the diagonal elements of the matrix are calculated from the full classical VAR (i.e., the diagonal elements are equal to those in the full VAR method, and the non – diagonal elements are set equal to zero).

Since  $\Sigma_\epsilon$  replaced by  $\hat{\Sigma}_\epsilon$ , we need only specify prior for VAR coefficient  $\theta$ . The Litterman prior assumes that the prior of  $\theta$  is  $\theta \sim N(\theta_0, V_0)$ .

$\theta_0 = 0$  (where the hyper-parameter  $\mu_1 = 0$ , which indicates a zero mean model) and nonzero prior covariance  $V_0 \neq 0$ . Note that although the choice of zero mean could lessen the risk of over-fitting, theoretically any value for  $\mu_1$  is possible.

To explain the Minnesota/Litterman prior for the covariance, note that the explanatory variables in the VAR in any equation can be divided into their lags of the dependent variable, lags of the other dependent variables, and finally, any exogenous variables, including the constant term. The elements corresponding to exogenous variables are set to infinity (i.e., no information about the exogenous variables is contained within the prior).

The remainder of  $V_0$  is then a diagonal matrix with its diagonal elements  $v_{ij}^l$  for  $l = 1, \dots, p$

$$v_{ij}^l = \begin{cases} \left( \frac{\lambda_1}{l^{\lambda_3}} \right)^2 & \text{for } (i = j) \\ \left( \frac{\lambda_1 \lambda_2 \sigma_i}{l^{\lambda_3} \sigma_i} \right)^2 & \text{for } (i \neq j) \end{cases} \quad (15)$$

where  $\sigma_i^2$  is the  $i$ -th diagonal element of  $\Sigma_\epsilon$ .

This prior setting simplifies the complicated choice of specifying all the elements of  $V_0$  down to choosing three scalars  $\lambda_1, \lambda_2$  and  $\lambda_3$ . The first two scalars  $\lambda_1$  and  $\lambda_2$  are overall tightness and relative cross – variable weight, respectively.  $\lambda_3$  captures the lag decay that, as lag length increases, coefficients are increasingly shrunk toward zero.

Note that changes in these hyper – parameter scalar values may lead to smaller (or larger) variances of coefficients, which is called tightening (or loosening) the prior. The exact choice of values for these three scalars depends on the empirical application, so researchers can make trials with different values for themselves. Litterman (1986) provides additional discussion of these choices.

Given this choice of prior, the posterior for takes the form

$$\theta \sim N(\bar{\theta}, \bar{V})$$

where

$$\bar{V} = \left[ V_0^{-1} + \left( \hat{\Sigma}_\epsilon^{-1} \otimes X'X \right) \right]^{-1} \tag{16}$$

and

$$\bar{\theta} = \bar{V} \left[ V_0^{-1} \theta_0 + \left( \hat{\Sigma}_\epsilon^{-1} \otimes X \right)' y \right] \tag{17}$$

A primary advantage of the Minnesota/Litterman prior is that it leads to simple posterior inference. The prior does not, however, provide a full Bayesian treatment of  $\Sigma_\epsilon$  as an unknown, so it ignores uncertainty in this parameter.

*Degrees of Freedom Correction*

This paper, we chosed:

The Prior specification: Hyper – parameters,

Coefficient Priors:  $\mu_1 = 0.$ , Residual Priors:  $\lambda_1 = 0.01.$ ;  $\lambda_2 = 0.99$  and  $\lambda_3 = 1.$

With the established parameters, we estimate the Bayesian VAR model.

**3.2.4 Diagnostic the Appropriateness of the Estimated Bayesian VAR**

We diagnostic the appropriateness of the estimated Bayesian VAR through:

*Stationary Test*

AR Roots Table the inverse roots of the characteristic AR polynomial; see Lütkepohl (1991). The estimated VAR is stable (stationary) if all roots have a modulus of less than one. If the VAR is not stable, specific results are not valid.

*Residual Tests*

To check the stability of the residuals, we use the autocorrelation diagram. Suppose the test results show that Q – statistics are insignificant at all lags, indicating no significant serial correlation in the residuals. In that case, the model’s residuals are white noise.

**3.2.5 Predict the Values of the Variables up to 2030**

After diagnosing the appropriateness, use the Bayesian VAR model to predict the values of the variables up to 2030.

**4 Research Results**

**4.1 Test for Stationarity**

Stationary test results (Table 1) show that the time series are stationary at the first difference, but VI stops at the second difference.

**4.2 Determine the Lag Intervals**

Base on Akaike information criterion (AIC), we determine the lag intervals for endogenous variables at 1.

**Table 1.** Kwiatkowski-Phillips-Schmidt-Shin test statistic

Time series	Asymptotic critical values*: 5% level	0.1460	Time series	Asymptotic critical values*: 5% level	0.1460
LM – Stat			LM – Stat		
D(US)	0.1499		D(FR)	0.5000	
D(GE)	0.1532		D(IT,2)	0.2343	
D(JA)	0.2037		D(NE,2)	0.5000	
D(CH,2)	0.3552		D(FR)	0.5000	
D(UK,2)	0.3380		D(VI,2)	0.1983	

Source: Analytical results from Eviews, [2022](#)

### 4.3 Diagnostic the Appropriateness of the Estimated Bayesian VAR

#### 4.3.1 Stationary Test

The test results show that all modulus values are less than 1, VAR satisfies the stability condition (Table 2).

**Table 2.** Stationary test

Roots of Characteristic Polynomial	
Endogenous variables: D(US) D(GE)	
D(JA) D(CH,2) D(UK,2) D(FR) D(IT,2)	
D(NE,2) D(SP) D(VI,2)	
Root	Modulus
-0.004588	0.9820
-0.004588	0.9820
-0.726307	0.9080
-0.726307	0.9080
0.403437	0.8591
0.403437	0.8591
-0.446040	0.8518
-0.446040	0.8518
0.011970	0.8317

(continued)

**Table 2.** (continued)

0.011970	0.8317
-0.686121	0.6861
0.473709	0.6109
0.473709	0.6109
0.592474	0.5959
0.592474	0.5959
-0.542744	0.5427
0.178066	0.5043
0.178066	0.5043
-0.382252	0.4441
-0.382252	0.4441
No root lies outside the unit circle	
VAR satisfies the stability condition	

Source: Analytical results from Eviews, 2022

**4.3.2 Residual Tests**

To check the stability of the residuals, we use the autocorrelation diagram. The test results show that Q – statistics are not significant at all lags, indicating no significant serial correlation in the residuals; the residuals are a stationary series (Fig. 2).

Autocorrelation	Partial Corelation		AC	PAC	Q-stat	Prob
		1	0.161	0.161	0.8657	0.353
		2	-0.310	-0.345	4.1548	0.125
		3	-0.150	-0.032	4.9562	0.175
		4	-0.161	-0.268	5.9121	0.206
		5	0.066	0.106	6.0816	0.298
		6	0.006	-0.222	6.0832	0.414
		7	-0.070	0.001	6.2901	0.506
		8	0.135	0.057	7.0821	0.528
		9	0.017	-0.060	7.0956	0.627
		10	-0.061	-0.009	7.2750	0.699
		11	-0.036	-0.063	7.3401	0.711
		12	0.035	0.115	7.4040	0.830

**Fig. 2.** Residual autocorrelation diagram. Source: Analytical results from Eviews, 2022

**4.4 Predict the Values of the Variables up to 2030**

Table 3 shows the forecast of textile export market share in the period 2020 – 2030 for Vietnam, the USA, Germany, Japan, China, England, France, Italy, the Netherlands, and

**Table 3.** Market share forecast

YEAR	CH_F	FR_F	GE_F	IT_F	JA_F	NE_F	SP_F	UK_F	US_F	VI_F
2020	9.206	3.503	5.577	3.146	3.369	1.997	2.908	3.676	14.32	3.155
2021	8.433	3.559	5.896	2.982	3.039	2.169	2.838	2.846	15.19	3.267
2022	7.855	3.427	6.493	2.886	3.966	2.141	3.005	2.376	15.05	3.454
2023	9.077	2.759	5.297	2.538	4.430	1.660	2.725	1.844	16.18	3.449
2024	9.354	2.850	4.997	2.585	4.102	1.548	2.711	1.541	15.92	3.304
2025	9.932	3.235	5.295	2.828	3.070	1.637	2.954	1.267	15.04	3.488
2026	9.098	3.237	5.822	2.752	3.288	1.795	3.055	0.7346	15.25	3.706
2027	9.880	2.711	5.118	2.492	3.878	1.386	2.978	0.4814	16.04	3.659
2028	11.210	2.418	3.973	2.372	3.355	1.006	2.803	0.009	16.93	3.584
2029	10.010	2.855	4.609	2.581	2.956	1.283	3.023	-0.5133	16.29	3.704
2030	9.752	2.893	4.969	2.615	3.191	1.352	3.221	-0.8748	15.84	3.938

Source: Analytical results from Eviews, [2022](#)

Spain. The forecast results show that the export market shares of three countries, Spain, the US, and Vietnam, tend to increase slightly and stabilize. The other seven countries, including China, France, Germany, Italy, Japan, the Netherlands, and England, tend to decrease.

For Vietnam, with the positive epidemic prevalence in 2020 of the Government, the Government has maintained that dual and both anti – epidemic and economic development has brought positive results. According to the General Statistics Office of Vietnam (Ha, [2021a](#), [2021b](#)), the world economy witnessed the most severe recession in history, the growth of significant economies decreased profoundly due to the harmful effects of the Covid – 19 epidemic, but Vietnam's economy still maintained growth with the growth rate of GDP estimated to reach 2.91%, bringing Vietnam to become a fourth largest economy in Southeast Asia (after Indonesia 1,088.8 billion USD; Thailand 509.2 billion USD and Philippines 367.4 billion USD).

However, 2021 will see the outbreak and collapse of many health systems in many countries due to the Delta mutation. This variant causes more infections and spreads faster than previous forms of SARS – NCOV – 2 (Nguyễn, N. D, [2022](#); Nguyễn, N. D., & Nguyễn, Y., [2022](#)). Moreover, this outbreak took a heavy toll in Vietnam between May and October 2021 with widespread outbreaks. In industrial parks, the epidemic spread rapidly, especially in industrial zones with textile factories with tens of thousands of workers, causing a wave of workers' migration from industrial zones to the countryside. The consequences of this wave of migration caused a series of orders from textile companies to move to other production factories in the region and other countries around the world.

The results in [Table 3](#) show that the market share of Vietnam's textile and garment exports still tends to increase gradually in the years 2022–2030. This result is similar to the study by Wei et al. ([2021](#)), which shows that the pandemic can increase the export

positions of countries; if the disease is adequately controlled, it will be a good opportunity for exports during this period, and vice versa.

## 5 Conclusion

Using the VAR Bayesian model and the data for the export market share of textiles collected from the World Bank, the author's team has forecast the export market share of textiles in the period 2020 – 2030 in 10 countries with the most prominent part (data in 2019). The research results show that the COVID – 19 pandemic has affected the export market share of textiles and garments. The forecast results show that the export market shares of three countries, Spain, the US, and Vietnam, tend to increase slightly and stabilize. The other seven countries, including China, France, Germany, Italy, Japan, the Netherlands, and England, tend to decrease.

The research results show that Vietnam's textile and garment export market share will continue to increase in the coming years. The COVID – 19 pandemic is the leading cause affecting the export value of Vietnam's textiles and garments. The findings show that the Vietnamese government needs to quickly implement measures to deal with the pandemic, take advantage of export opportunities during the pandemic, and bring many economic benefits.

The current perspective on the current epidemic situation and the application of instant responding measures from time to time to control the epidemic in parallel with economic development is essential. Conduct harmonization between priority prevention and control of epidemics and socio – economic development; only make ways of stretching and blocking when it is essential and within the scope of consistency and enforcement of the goal to continue to maintain macroeconomic stability; control inflation; ensure large balances of the economy; Implementing harmony, successful double target prevention and control of COVID – 19 epidemic, has promoted socio – economic development.

Textile and garment exporters can use this research result to proactively better prepare for production and prepare human resources after the wave of workers leaving industrial zones to avoid the end of the epidemic. Better preparation will help textile and garment exporters to be more efficient in their production and export activities and maintain their market share in textile exports.

Vietnam needs to focus on diversifying export markets, actively seeking new markets, and effectively exploiting opportunities from FTAs. Monitoring developments, analyzing and forecasting the impact of the Covid – 19 pandemic on the value of Vietnam's exports will contribute to updating the economic growth scenario and helping Vietnam to remove difficulties and obstacles in the export process. Ensure convenient and safe customs clearance to stabilize and develop the economy during the Covid – 19 pandemic.

## References



Barua, S., Kar, D., Mahbub, F.B.: Risks and their management in ready-made garment industry: evidence from the world's second largest exporting nation. *J. Bus. Manag.* **24**, 75–103 (2018)



- Do, K.D.: Evaluating the competitiveness of the vietnam textile and garment industry. *J. Int. Bus. Manag.* **4**(10), 1–13 (2021). <https://doi.org/10.37227/JIBM-2021-08-1176>
- EViews Help: <http://www.eviews.com/help/helpintro.html#page/content/preface.html> (n.d.). Retrieved 7 Nov 2021
- Goto, K., Natsuda, K., Thoburn, J.: Meeting the challenge of China: the Vietnamese garment industry in the post MFA era. *Global Netw.* **11**(3), 355–379 (2011). <https://doi.org/10.1111/j.1471-0374.2011.00330.x>
- Ha, T.: Kinh tế Việt Nam 2020: Một năm tăng trưởng đầy bản lĩnh. General Statistics Office of Vietnam. <https://www.gso.gov.vn/du-lieu-va-so-lieu-thong-ke/2021a/01/kinh-te-viet-nam-2020-mot-nam-tang-truong-day-ban-linh/> (2021a)
- Ha, T.: Triển vọng kinh doanh của ngành dệt may trong năm 2020. General Statistics Office of Vietnam. <https://www.gso.gov.vn/du-lieu-va-so-lieu-thong-ke/2021b/07/trien-vong-kinh-doanh-cua-nganh-det-may-trong-nam-2020/> (2021b)
- Kreinovich, V., Thach, N.N., Trung, N.D., Thanh, D.V. (eds.): *Beyond Traditional Probabilistic Methods in Economics*. Springer International Publishing, Cham (2019)
- Megersa, K.: Structure of the Global Ready-Made Garment Sector. <https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/14705> (2019)
- Ministry of Industry and Trade: Báo cáo Xuất nhập khẩu Việt Nam 2020: Kho thông tin hữu ích cho doanh nghiệp. <https://congthuong.vn/bao-cao-xuat-nhap-khau-viet-nam-2020-kho-thong-tin-huu-ich-cho-doanh-nghiep-155486.html> (2021)
- Nguyễn, N.D.: Export and application of SARIMA model for forecasting the value of Vietnam export during Covid-19. *VNUHCM J. Econ. Bus. Law* **6**(2), 2832–2839 (2022). <https://doi.org/10.32508/stdjelm.v6i2.891> <http://stdjelm.scienceandtechnology.com.vn/index.php/stdjelm/article/view/891>
- Nguyen, T.N., Kosheleva, O., Kreinovich, V., Nguyen, H.P.: Blockchains beyond bitcoin: towards optimal level of decentralization in storing financial data. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *Beyond Traditional Probabilistic Methods in Economics*, pp. 163–167. Springer International Publishing, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_12](https://doi.org/10.1007/978-3-030-04200-4_12)
- Nguyễn, N.D., Nguyễn, Y.: Development of supporting industries towards production and export autonomy in Ho Chi Minh City. *VNUHCM J. Econ. Bus. Law* **6**(3), 3386–3395 (2022). <https://doi.org/10.32508/stdjelm.v6i3.1031>; <http://stdjelm.scienceandtechnology.com.vn/index.php/stdjelm/article/view/1031>
- Textiles and Clothing Exports by country & region US\$000 2015 | WITS Data. (n.d.). [https://wits.worldbank.org/CountryProfile/en/Country/WLD/Year/2015/TradeFlow/Export/Partner/all/Product/50-63\\_TextCloth](https://wits.worldbank.org/CountryProfile/en/Country/WLD/Year/2015/TradeFlow/Export/Partner/all/Product/50-63_TextCloth). Retrieved 29 Oct 2021
- Thach, N.N., Kreinovich, V., Trung, N.D. (eds.): *Data Science for Financial Econometrics*, vol. 898. Springer International Publishing, Cham (2021). <https://doi.org/10.1007/978-3-030-48853-6>
- Thach, N.N., Kreinovich, V., Ha, D.T., Trung, N.D. (eds.): *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty, and Related Topics*, vol. 427. Springer International Publishing, Cham (2022). <https://doi.org/10.1007/978-3-030-98689-6>
- Wei, P., Jin, C., Xu, C.: The influence of the COVID-19 pandemic on the imports and exports in China, Japan, and South Korea. *Front. Public Health* **9**, 682693 (2021). <https://doi.org/10.3389/fpubh.2021.682693>
- wits.worldbank (n.d.) Textiles and Clothing Exports by country & region. [https://wits.worldbank.org/CountryProfile/en/Country/WLD/Year/2019/TradeFlow/Export/Partner/all/Product/50-63\\_TextCloth](https://wits.worldbank.org/CountryProfile/en/Country/WLD/Year/2019/TradeFlow/Export/Partner/all/Product/50-63_TextCloth). Retrieved 21 Sep 2021



# The Impact of Global Value Chain Integration on Export: Evidence from Vietnam

Nguyen Thi Ngoc Diep<sup>1,2</sup>(✉) , Tran Quang Canh<sup>3</sup>(✉) , and Nguyen Ngoc Thach<sup>4</sup>

<sup>1</sup> University of Economics and Law, Ho Chi Minh City, Vietnam  
diepntn@uel.edu.vn

<sup>2</sup> Vietnam National University, Ho Chi Minh City, Vietnam

<sup>3</sup> Ho Chi Minh City University of Economics and Finance, Ho Chi Minh City, Vietnam  
canhtq@uef.edu.vn

<sup>4</sup> Banking University of Ho Chi Minh City, 36 Ton That Dam Street, District 1, Ho Chi Minh City, Vietnam  
thachnn@buh.edu.vn

**Abstract.** The article studied the impact of the global value chain integration level on Vietnam's exports from 1990 – 2018. The study used a Bayesian linear regression analysis approach to assess the impact of foreign-origin added value (FVA), domestic added value (DVA), and domestic value-added export (DVX) of other countries in comparison with the value of Vietnam's exports (EXPORT). The findings showed positive relationships between foreign – origin added value (FVA), domestic added value (DVA), domestic value-added export of other countries, and export value of Vietnam. The more Vietnam is integrated into the global value chain, the more export value increases.

**Keywords:** Global Value Chain (GVC) · export value · Bayesian model, Vietnam's exports

## 1 Introduction

In the strategic period of the past ten years, from 2011 to 2020, Vietnam formed some critical industries of the economy. Those such as oil and gas exploitation and processing; electronics, telecommunications, information technology; metallurgy, iron, and steel; cement and construction materials; textiles, footwear; mechanical engineering, automotive, and motorcycles, which have created an essential foundation for long-term growth and promotes the process of modernization and industrialization. In particular, the industry is the sector with the highest growth rate among the national economic sectors, with a contribution of approximately 30% to GDP, which becomes the primary export sector of the country and contributes to bringing Vietnam from the 50th position (2010) to the 22nd place (2019) among the world's largest exporting countries.

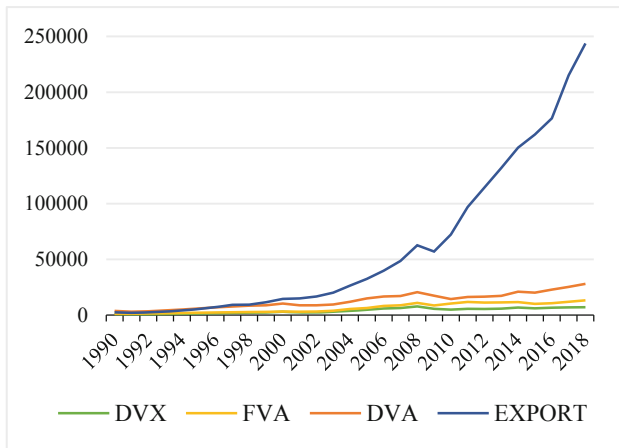
According to the World Development Report of the World Bank (2020), most countries in the world participate in global value chains related to cross – border transactions

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 440–448, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_31](https://doi.org/10.1007/978-3-031-35763-3_31)

and transactions with buyers in worldwide and international companies. Globally, up to two – thirds of world trade occurs through a global value chain, where production passes through at least one border before reaching the final assembly stage. In Vietnam, the trade balance in goods in the 2016 – 2020 period continuously had a trade surplus from 1.6 billion USD in 2016 to nearly 20 billion USD in 2020. In 2020, the value of exports reached US \$282.63 billion; the value of imports reached US \$262.69 billion in 2020; and the trade surplus reached US \$19.94 billion (Firmani, n.d.). According to preliminary data from the Government of Vietnam, the manufacturing sector contributed significantly to the expansion of the trade balance, with an industrial trade surplus exceeding 10 billion USD for the first time in 2019. Vietnam is a net importer of services, with imports at \$18.2 billion and exports at \$14.9 billion in 2018.



**Fig. 1.** Global value chain value (including DVX, FVA, and DVA) and export value of Viet Nam. Source: *World Integrated Trade Solution (WITS) | Data on Export, Import, Tariff, NTM*, n.d.

The Minister of Industry and Trade of Vietnam said that a motive in the production and export of industrial products of Vietnam is still mainly driven by the Foreign Direct Investment (FDI) sector, accounting for approximately 70% of the country's total export turnover. Production and export dynamics are affected by FDI, mainly because of weak linkages between FDI and domestic enterprises in the supply chain. It also shows the limited competitiveness of domestic enterprises when participating in global value chains (Fig. 1). This article was designed to study the influence of the components constituting the level of global value chain integration (including DVX, FVA, and DVA) on Vietnam's export value in the period 1990 – 2018 to recognize and evaluate the impact of these components on the export value of Vietnam. Hence, our study recommended improving the Vietnamese economy's global value chain integration level.

Our study includes the literature review in part 2, the research method in part 3, the research results in part 4, and the conclusion and policy implications in part 5.

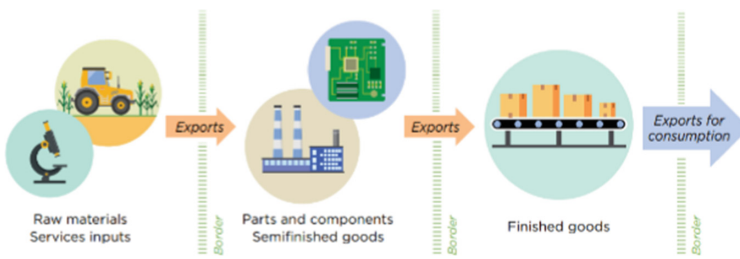
## 2 Literature Review

### 2.1 Value Chain and Global Value Chain (GVC)

The value chain involves all essential activities to produce a product or service from inputs, which go through different stages of production (i.e., the combination and transformation of input material and various production services), bring that product to the end consumer, and performs the service after use.

Porter, (2001) separated the supply chain stages (such as inbound logistics, operations, outbound logistics, marketing, sales, and after-sales services) based on his research. A business gets its job done by converting inputs like production processes, warehousing, quality, and continuous improvement, as well as the supporting services it needs (e.g., strategic planning, human resources, technology development, procurement).

According to the (World Bank, 2020), the Global Value Chain (GVC) refers to the production division between countries. Countries participating in the value chain were positively correlated to exporting inputs to other countries to produce their exports. In contrast, they were negatively related to import inputs to process and continue exporting (Fig. 2).



**Fig. 2.** Global Value Chain. Ource: World Bank, 2020

According to United Nations Conference on Trade and Development (2013), the primary global value chain indicators include:

**Foreign Origin Value Added (FVA):** The FVA indicates which part of a country's total exports includes inputs from other countries. The FVA share is the proportion of a country's exports that do not add to the country's GDP.

- **Domestic value added (DVA):** The DVA is the part of exports generated in the country, which means that exports contribute to GDP.
- **Domestic value added integrates into other countries' exports (DVX):** This index indicates the extent to which a country's exports are used as inputs for other countries' exports.
- **The global value chain participation index** determines the percentage of participation in exports by a country in a multi – stage commercialization process.

Although the level of exports used by other countries may be less relevant to policymakers since it does not change the DVA contribution of trade, the participation rate is a valuable indicator of how much a country's exports integrate into the international production network.

## 2.2 Export and Global Value Chain (GVC)

A country's exports can be divided into domestically produced added value and imported foreign added value which integrates into that country's exported goods and services. Moreover, exports to foreign markets can be regarded as the final consumer goods or inputs for producing goods to export to third countries (or back to the original country). Analysis of the Global value chain encompasses foreign value added in exports (upstream view) and export value added with the integration of third-country exports (downstream view). Today's production processes are structured in several stages and occur in many countries. To produce the final product, companies seek input sourcing from many suppliers; in many cases, these suppliers are located overseas (Nguyễn, N. D., 2022; Nguyễn, N. D., & Nguyễn, Y., 2022). The added value at each stage of the production process and the product can cross the borders several times before being consumed in the last place. The motive of efficiency and cost consideration is considered behind decisions to use foreign inputs or locate production stages, including final assembly abroad. For example, smartphone research and development may occur in an economy with specific competitive advantages. At the same time, the final product is assembled where labor costs are relatively low, and the components are supplied by the countries that specialize in producing them.

Each country involved in the manufacturing process contributes to the total added value of the final product despite different rates. The global value chain has proven to be an important channel for technology transfer between countries.

The opportunity to transfer know-how, technology, and process innovation through participation in the global value chain is enormous. Enterprises can access new technology that integrates into import inputs and benefit from new types of intermediate goods by expanding inputs.

## 3 Methodology

### 3.1 Model Research

Bayesian analysis is a powerful analytical tool for statistical modeling, interpretation of results, and data prediction. In Bayesian analysis, the estimation accuracy is not limited by sample size or by limitations such as autocorrelation, endogenous, or heteroscedasticity encountered by frequency methods (Canh & Hoai, 2022; Kreinovich et al., 2019; Nguyen T.N et al., 2019; Thach N.N et al., 2021; Thach N.N et al., 2022). Since the data collected is limited, there is only data from 2010 to 2020 on the Vietnam General Statistics Office website. Therefore, using Bayesian regression is appropriate.

The study data was analyzed using BayES 2.4 software. The study uses the Bayesian linear regression model with specific research model as follows:

$$\ln EXPORT = constant + \beta_1 \ln DVX + \beta_2 \ln FVA + \beta_3 \ln DVA + \varepsilon \text{ with } \varepsilon \sim N\left(0, \frac{1}{\tau}\right) \quad (1)$$

In which:

N: Number of observations.

lnDVX: The natural logarithm of indirect value added, domestic value added integrating into exports of other countries.

lnFVA: The natural logarithm of value added originating from abroad.

lnDVA: The natural logarithm of value added originating from domestic.

lnEXPORT: The natural logarithm of Vietnam’s export value.

$\beta$ : Regression coefficient of the model.

$\varepsilon$ : Residuals.

$\tau$ : is the precision of the error term:  $\sigma_\varepsilon^2 = \frac{1}{\tau}$ .

The predetermined values of the research model were set as the default of the software (Emvalomatis, 2020), specifically as in Table 1.

**Table 1.** The predetermined values of the research model

Parameters	Probability density function	Predefined hyperparameters
$\beta$	$p(\beta) = \frac{ P ^{1/2}}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(\beta - m)'P(\beta - m)\right\}$ (2)	$m = 0_k, P = 0.001 * I_k$
$\tau$	$P(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau - 1} e^{-\tau b_\tau}$ (3)	$a_\tau = 0.001, \beta_\tau = 0.001$

Source: Default of the software, 2022

The optional arguments for the model are:

Gibbs parameters:

Chains: Number of chains to run in parallel (positive integer); the default value is 1.

Burnin: Number of burn-in draws per chain (positive integer); the default value is 10000.

Draws: Number of retained draws per chain (positive integer); the default value is 20000.

Thin: Value of the thinning parameter (positive integer); the default value is 1.

Seed: Value of the seed for the random – number generator (positive integer); the default value is 42.

Hyperparameters:

m: Mean vector of the prior for  $\beta$  ( $K \times 1$  vector); the default value is 0K.

Precision matrix of the prior for  $\beta$  ( $K \times K$  symmetric and positive-definite matrix); the default value is  $0.001 \cdot IK$ .

a\_tau: Shape parameter of the prior for  $\tau$  (positive number); the default value is 0.001.

b\_tau: Rate parameter of the prior for  $\tau$  (positive number); the default value is 0.001.

Dataset and log-marginal likelihood.

Dataset: The id value of the dataset that will be used for estimation; the default value is the first dataset in memory (in alphabetical order).

logML\_CJ: Boolean indicating whether the Chib (1995) and Chib & Jeliazkov (2001) approximation to the log-marginal likelihood should be calculated (true/false); the default value is false.

The regression model is satisfactory when the following tests are satisfied Emvalomatis, (2020): the value of regression coefficients is statistically significant when the regression coefficient is different from zero in the 90% confidence interval, the tau values of the variables are positive, and the Inefficiency factor  $> 1$ .

To assess the convergence of the MCMC series, the authors used the trace plots with the requirement that the MCMC errors of the post – test means are decimals less than 0.05.

Plots of the simulated probability series (draw index) with rapid fluctuations do not show anomalous sequences. The correlation plot between the lags of the draws for  $\tau$  and the variables rapidly decreases to oscillate around zero, which indicates that they are not autocorrelated. The frequency charts and density histogram of the draws are smooth, showing that the sampler has no abnormal phenomenon for any amount drawn in specific regions of the sample space.

### 3.2 Data

Research data on DVX, FVA, and DVA were collected from The World Trade Organization (WTO), and EXPORT was collected from Vietnam General Statistics Office. Mainly, DVX refers to indirect value-added, domestic value-added, which is integrated into export goods of other countries. FVA refers to value added originating from abroad. DVA means value added originating from domestic. EXPORT refers to the export value of Vietnam.

## 4 Research Results

### Model Evaluation

The results of the stationarity test showed that all variables were stationary at the first difference level. The regression result for the first difference level of the variables is presented in Table 2. The first column of the table contains the names of the variables to which each parameter is linked. The second column (Mean) contains the posterior mean of each parameter, and the third (Median) and fourth columns (Sd.dev) are the median and standard deviation, respectively. The last two columns (5% and 95%) give the value of the first and last points of the 90% confidence intervals.

The regression results showed that the  $\ln DVX$  variable was not statistically significant except for the regression coefficient of the first difference. The values of the remaining regression coefficients differed from zero in the 90% confidence interval, and the tau value of the variables were positive and Inefficiency factor  $> 1$  (see Table 2).

Thus, there is evidence to conclude that the regression coefficients (except  $\ln DVX$ ) are statistically significant.

Simulation results showed that the MCMC errors of the posterior means were decimals less than 0.05. According to Emvalomatis (2020), the authors used trace plots to evaluate the convergence of MCMC chains.

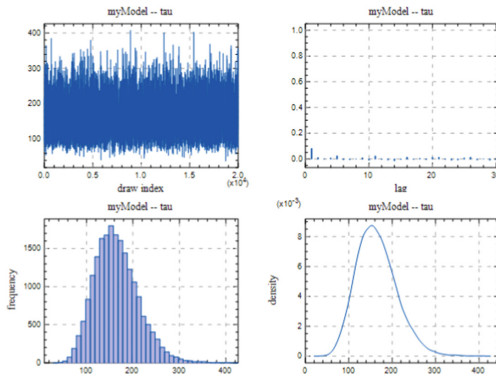
Figure 1 showed that the plot of the draw index sequence fluctuated rapidly and did not show anomalous sequences.

**Table 2.** Regression results for dependent variable lnEXPORT

Variables	Mean	Sd.dev	MCMC sd. Error	5%	95%	Inefficiency factor
constant	0.1086	0.193	0.0001	0.0768	0.1402	1.0107
D(lnDVA)	0.4803	0.1399	0.0010	0.2482	0.7089	1.0392
D(lnFVA)	0.2460	0.1436	0.0010	0.0107	0.4793	1.0355
tau	164.155	46.2133	0.3632	95.5678	246.362	1.2354

Source: Results of the authors’ analysis, 2022

The correlation chart between the lagging of the draws for t and the variables indicated no autocorrelation. Frequency histograms and density charts of the draws were smooth, showing that the sampler had no abnormalities for any significant draw in specific regions of the sample space (Fig. 3).



**Fig. 3.** Trace chart of variables. Source: Results of the authors’ analysis, 2022.

Simulation results also show that the MCMC errors (see Table 2) of the posterior means were decimals less than 0.05. The analysis results of the trace plot and the value of MCMC errors showed that the MCMC chain converges, so it was possible to trust the results of Bayes’ inference.

## 5 Conclusion

The results showed that there is a positive impact between the level of integration into the global value chains and Vietnam’s export value. It shows that the level of integration into global value chain of Vietnam needs to be further increased, which will contribute to improving the value of exports, enhancing commercial position regionally and internationally. Findings showed that Vietnam needs to focus on the domestic value added with its integration into exports of other countries (DVX), which positively impacts on the



export value of Vietnam and promotes Vietnam to gradually become an upstream country in the global value chain (i.e., the exporter of goods to other countries for processing).

The results positively impacted the level of integration into the global value chains and Vietnam's export value. It shows that the level of integration into the global value chain of Vietnam needs to be further increased, which will contribute to improving the value of exports and enhancing its commercial position regionally and internationally. It was determined that Vietnam needs to emphasize its domestic value added by integrating into other countries' exports (DVX). It positively impacts Vietnamese exports and helps Vietnam gradually become an upstream country in global value chains (i.e., exporting goods to another country for processing).

It is necessary to improve the domestic value added to Vietnamese exports to create high export value. In addition, it should be recognized that the positive impact between the level of integration into the global value chain and the export value of Vietnam is low and needs to be commensurate with the export value of Vietnam. In fact, according to the data of the Vietnam Chamber of Commerce and Industry, the participation of Vietnamese enterprises in the global value chain was meager compared to economies of similar size in Southeast Asia. Specifically, only 36% of Vietnamese enterprises participated in the production network, including direct and indirect exports, while this percentage in Malaysia and Thailand was 60%.

Vietnamese enterprises were fragmented and were less likely to benefit from the spillover effect of foreign investment, technology transfer, knowledge transfer, and improved production capacity. The main reason for this situation is that Vietnam has only about 4% of enterprises with sufficient competitiveness that can participate in the global supply chain.

In Vietnam, domestic enterprises are almost unrelated to global value chains like foreign direct investment enterprises. It is a challenging requirement for Vietnamese enterprises in the current period, making it difficult to break into the global supply chain. Therefore, to improve competitiveness for domestic enterprises, there should be close coordination and joint hands of stakeholders, including the Government, ministries, industry, and localities, in defining strategic priorities, creating policy frameworks, and improving the investment and business environment. Additionally, the role of investors and enterprises must be enhanced in innovation, capacity building, and seeking opportunities from new trends. Vietnam's ability to participate in the global value chain also emerges from the Government's firm commitments to improve the business environment through improving key indicators, such as market entry, access to electricity, and intellectual property. Hence, policymakers must be flexible and well – planned so that Vietnam is always an attractive destination for foreign investment, ready to supply highly skilled labor for the economy to catch up with the global market's growth momentum and profoundly integrate into the global value chain.

## References

- Canh, T.Q., Hoai, P.T.D.: Factors affecting gdp per capita—Apply Bayesian analysis. *VNUHCM J. Econ. Bus. Law* 6(2), 2 (2022). <https://doi.org/10.32508/stdjelm.v6i2.961>
- Chib, S.: Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc.* 90(432), 1313–1321 (1995)

- Chib, S., Jeliazkov, I.: Marginal likelihood from the Metropolis-Hastings output. *J. Am. Stat. Assoc.* **96**(453), 270–281 (2001)
- Emvalomatis, G.: Basic linear model. In: *User's Guide BayES<sup>TM</sup> Bayesian Econometrics Software*. [http://bayeconsoft.com/html\\_documentationse25.html#x31-440001](http://bayeconsoft.com/html_documentationse25.html#x31-440001) (2020)
- Firmani, M.: PX Web. General Statistics Office of Vietnam. <https://www.gso.gov.vn/en/px-web/> (n.d.). Retrieved 5 Nov 2022
- Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.): *Beyond Traditional Probabilistic Methods in Economics*. Springer International Publishing, Cham (2019)
- Nguyễn, N.D.: Export and application of SARIMA model for forecasting the value of Vietnam export during Covid-19. *VNUHCM J. Econ. Bus. Law* **6**(2), 2832–2839 (2022). <https://doi.org/10.32508/stdjelm.v6i2.891>; <http://stdjelm.scienceandtechnology.com.vn/index.php/stdjelm/article/view/891>
- Nguyễn, N.D., Nguyễn, Y.: Development of supporting industries towards production and export autonomy in Ho Chi Minh City. *VNUHCM J. Econ. Bus. Law* **6**(3), 3386–3395 (2022). <https://doi.org/10.32508/stdjelm.v6i3.1031>; <http://stdjelm.scienceandtechnology.com.vn/index.php/stdjelm/article/view/1031>
- Nguyen, T.N., Kosheleva, O., Kreinovich, V., Nguyen, H.P.: Blockchains beyond bitcoin: towards optimal level of decentralization in storing financial data. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *ECONVN 2019*. SCI, vol. 809, pp. 163–167. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_12](https://doi.org/10.1007/978-3-030-04200-4_12)
- Porter, M.E.: The value chain and competitive advantage. *Understanding Bus. Proc.* **2**, 50–66 (2001)
- Thach, N.N., Kreinovich, V., Trung, N.D. (eds.): *Data Science for Financial Econometrics*. SCI, vol. 898. Springer, Cham (2021). <https://doi.org/10.1007/978-3-030-48853-6>
- Thach, N.N., Kreinovich, V., Ha, D.T., Trung, N.D. (eds.): *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty, and Related Topics*, vol. 427. Springer International Publishing, Cham (2022). <https://doi.org/10.1007/978-3-030-98689-6>
- United Nations Conference on Trade and Development: *World Investment Report 2013: Global Value Chains – Investment and Trade for Development*. UN (2013). <https://doi.org/10.18356/a3836fcc-en>
- World Bank: *World Development Report 2020: Trading for Development in the Age of Global Value Chains*. World Bank, Washington, DC (2020). <https://doi.org/10.1596/978-1-4648-1457-0>
- World Integrated Trade Solution (WITS) | Data on Export, Import, Tariff, NTM. <https://wits.worldbank.org/> (n.d.). Retrieved 5 Nov 2022



# A Hybrid Model Based on ARIMA and Artificial Neural Network to Forecast Consumer Price Index: The Case of Vietnam

Thi Thanh Huyen Le<sup>1</sup> and Tien Nhat Nguyen<sup>2</sup>(✉)

<sup>1</sup> Ho Chi Minh University of Banking, Ho Chi Minh, Vietnam  
huyenl1tt@hub.edu.vn

<sup>2</sup> Hue College of Economics, Hue University, Hue, Vietnam  
ntnhat@hce.edu.vn

**Abstract.** While the Autoregressive Integrated Moving Average (ARIMA) model has been dominantly used to capture a linear component of time series data in the field of economic forecast for years, the Artificial Neural Networks (ANNs) increasingly are applying to explore tough challenge due to an existence of both linear and nonlinear patterns in a certain time series dataset. Regarding the time series forecasting, most studies have applied only ARIMA model or ANNs to produce multiple-step prediction with an insufficiently reasonable accuracy. That's why this paper suggests a hybrid model with the advantages of either ARIMA or ANNs to analyze the linear and nonlinear relationships in Vietnam CPI time series from January 1995 to July 2022. The result shows that an effectiveness in the multiple-step prediction of the hybrid model is more precise in comparison with ARIMA model and ANNs.

**Keywords:** ARIMA · Artificial Neural Networks (ANNs) · Nonlinear autoregressive neural network (NAR) · hybrid model · time series forecast · Vietnam · Consumer price index (CPI) · inflation

## 1 Introduction

In a field research of time series forecasting, (Box, 2013) developed the integrated autoregressive moving average (ARIMA) methodology which have been already indicated the restriction of linearity by statisticians in several ways. That's why the robust versions of various ARIMA models have been evolved to function as a powerful tool as forming the empirical dependencies between successive and failures times (Walls & Bendell, 1987). Nonetheless, a major disadvantage of these models is a fundamental assumption with having an existence linear correlation among the time series values which unfortunately differs with roughly problems in practice typically considered as existing in a non-linear relationship. Therefore, alternative methods that thoroughly investigate the non-linear interaction need to be developed due to unsatisfactory results of the estimation of linear models to factual problems. One of radical alternatives is Artificial Neural Networks

(ANNs) model whose properties imbricate statistical approaches in considerable way (Bishop, 1995), with performing a statistical analysis mainly based on an availability of data. An evolution of applied artificial neural systems consisting of ANNs has allowed a development of forecasting future values and knowledge depended on support of decision, having the fact that their patterns of recognition are considered uncomplicated and can be employed to a wide range of practical fields (Robert R. Trippi, 1996) where traditionally statistical methods are dominant. A major advantage of neural networks adequately reflects in their resilience as building models of non-linear relationship between economic factors that adaptively depended on the characteristics signified in the data. Therefore, ANNs can be seen as the data-based approach which well matches to several empirical data sets with having not an availability of theoretical guidance used as a suggestion for a proper process of data generation. In the field of time series forecasting, numerous empirical studies in which there are varied research of large-scale data applied to produce economic forecasts have demonstrated an effectiveness of ANNs. In comparison with ARIMA, ANNs can be also considered as nonparametric techniques with having a fundamental similarity of modelling a proper enclosed illustration of time series data.

Several research has applied ANNs in a way of integrating the ANNs with the ARIMA models, so that they could take advantages of these two models to generate the most accurate results (Kohzadi et al., 1996). A concept of hybrid models which is a combination of ANNs and ARIMA model in which each model's unique feature properly integrates into an organic entity so as to seize varied patterns in the data have been demonstrated to improve an accuracy of forecasting (Bates & Granger, 1969). The result of forecasting was proven to be more effective and efficient as combining more than one forecasting model with the well-known M-competition problem (Makridakis et al., 1982) according to both theoretical and empirical findings ((Palm & Zellner, 1992); (Ginzburg & Horn, 1993); (Luxhøj et al., 1996)). Some hybrid models have been presented in forecasting research with the application of ANNs. For example, the model built based on radial basis function networks of the Box–Jenkins models (Wedding & Cios, 1996); the hybrid synthesis of multiple models (Sfetsos & Siriopoulos, 2004) aiming to analyze a conformity between an efficiency of model and clustering algorithms and neural network. In addition, the hybrid artificial intelligence model considered as a formation of the rule-based system and the neural networks techniques used to predict the daily trend of S&P 500 Index (Tsaih et al., 1998), the integrated model of neural network and fuzzy model applied to predict the exchange rate (Kodogiannis & Lolis, 2002); or the forecast of price trend in short term of Taiwan stock index which was produced by the neural network with training data from ARIMA outcomes (Wang & Leu, 1996). Some studies demonstrated that the hybrid model provided more accurate forecasts than either the ARIMA model or the ANNs (P. G. Zhang, 2003), reflecting in better values of MSE (Mean Square Error). The hybrid model with the combination of ARIMA and ANNs apparently improves the preciseness of forecasting in a way that the ARIMA would manipulate the linear correlation lying in the historical data, while the ANNs would examine the nonlinear interaction existing in the uncertain parts of data. Basically, the time series composes four parts of trend (T), cyclical alteration (C), seasonal change (S), and irregular fluctuation (I), thus there is a common pattern of

model to analyze the time series data: an additive model ( $TS = T + C + S + I$ ). Besides, the correlations dwelling in the time series data consist of linear part (L) and nonlinear part (N), hence the additive model is (L + N).

The priority objective of this study is to build a hybrid of ARIMA and ANNs models to forecast the GDP and CPI of Vietnam. The first reason why the hybrid model is the chosen approach for this paper is to overcome an issue of what appropriate models for a certain time series dataset that could have linear or nonlinear structures or both. Another reason is that there is no fitted model which could deal with every data sample in terms of time-series forecasting (Chatfield, 1988); (Jenkins, 1982), because of a complexity of problems in practice and a lack of ability to explore varied structures in an efficient way of a certain model. (G. P. Zhang et al., 2001), for instant, applied ARIMA models to evaluate an efficiency of outcomes produced by ANNs model in the field of time-series forecasting. (Makridakis et al., 1993) proposes that the hybrid model formed by integrating several various models is likely to produce more accurate results as forecasting in comparison with a single model regardless of a consideration of what the best model is. S. Makridakis also found that a popularity of M2-contest which demonstrated that an integration of varied models in forecast enhanced an accuracy of an outcome, significantly initiated the application of the hybrid model in the field of time-series forecasting. From the initial study of (Bates & Granger, 1969) to inclusive research of (Clemen, 1989), the field of time-series forecasting using the hybrid model has undergone a long period of development with a core idea of taking an advantage of each model's specification to explore all types of structure lying in the dataset. For example, (Wedding & Cios, 1996) proposed the hybrid model of radial basis function networks and the Box-Jenkins models, (Naftaly et al., 1994) introduced the hybrid model of plentiful feedforward neural networks.

The rest of the paper is presented as follows: the next section reviews the theoretical basis of ARIMA and ANNs approaches to time-series forecasting and the hybrid model. Section 3 presents the process of selecting and designing the hybrid model as well as the data description. Then Sect. 4 reports empirical results from two datasets. Finally, Sect. 5 proposes the concluding remarks.

## 2 Methodology

### ARIMA Model

ARIMA model has a basic assumption about the linear correlation between the predicted value of dependent variable and the combination of historical observations of independent variables and random errors which can be illustrated as follows:

$$y_t = \vartheta_0 + \vartheta_1 y_{t-1} + \vartheta_2 y_{t-2} + \Lambda + \vartheta_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \Lambda - \theta_q \varepsilon_{t-q} \quad (1)$$

where  $y_t$  and  $\varepsilon_t$  are the actual value and random error at time  $t$ , while  $\vartheta_i$  ( $i = 1, 2, \dots, p$ ) and  $\theta_j$  ( $j = 1, 2, \dots, q$ ) are model parameters; and  $p$  and  $q$  are integers considered as orders of autoregressive and moving average polynomials. This model assumes that  $\varepsilon_t$  is

distributed in an independent and identical way with having a mean zero and a constant variance ( $\sigma^2$ ). As  $q = 0$ , Eq. (1) is an autoregressive (AR) model with  $p$  orders. As  $p = 0$ , Eq. (1) is a moving average (MA) model with  $q$  orders. The seasonal model is expressed as ARIMA ( $p, d, q$ ) ( $P, D, Q$ ). The method of forming ARIMA models evolved by Box and Jenkins (1970) allowed to significantly advance the time series analysis and forecasting application, which was dominantly applied in various fields of time series forecasting for years. In addition, Box and Jenkins (1970) developed a process of ARIMA modelling including (i) model identification, (ii) parameter estimation and (iii) diagnostic test, with a fundamental assumption that a time series produced by ARIMA process is likely to contain theoretical autocorrelation patterns. It also suggested that the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the sample data should be used to select the value of  $p$  and  $q$  for ARIMA model.

An important requirement of the model identification stage is that the time series data should be in a form of stationary, meaning that the mean and autocorrelation pattern are unchanged regardless of timing point along the period. If the time series shows the trend and heteroscedasticity, it should be applied to a differencing transformation to terminate the trend and fix the variance. Secondly, the aim of parameters estimation stage is to minimize overall values of errors by utilizing a process of nonlinear optimization. Finally, some diagnostic tests are implemented to check a sufficiency of model by using the goodness of the error term as a main criterion.

**Artificial Neural Networks (ANNs) Model**

The artificial neural network has a proven capability for mapping intricate nonlinear relationships while it unnecessarily depends on assumptions of acquiring an essence of the relationship. A well-known ANNs model is the feed-forward network having input layer and single hidden layer that are formed in nonlinear s-shaped functions, and output layer represented in linear transfer function. For more details, the output ( $y_t$ ) link to the inputs ( $y_{t-1}; y_{t-2}; \dots; y_{t-p}$ ) as follows:

$$y_t = \alpha_0 + \sum_{j=1} \alpha_j g \left( \beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t \tag{2}$$

$\alpha_j$  ( $j = 0, 1, 2, \dots, q$ ) and  $\beta_{ij}$  ( $i = 0, 1, 2, \dots, p; j = 1$ ) are the connection weights in which  $p$  is the number of input nodes and the number of hidden nodes is 1. This one-output-node network can afford to estimate an arbitrary function with a condition that the number of hidden nodes is set up to be sufficient.

The hidden layer is usually in the form of logistic function as follows:

$$g(x) = \frac{1}{1 + \exp(-x)} \tag{3}$$

Basically, the ANNs expresses a nonlinear function connecting the past observations ( $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ ) to the future value  $y_t$  as follows:

$$y_t = f(y_{t-1}, y_{t-2}, \Lambda, y_{t-p}, w) + \varepsilon_t \tag{4}$$

where  $w$  is a vector of whole parameters;  $f$  is a function formed by the network structure and connection weights, hence ANNs can be seen as a nonlinear autoregressive model.

In the process of designing neural network, there are no systematic obligations as selecting appropriate number of hidden layers  $q$ , it just depends on the characteristics of data of research. Besides, the good choice of number of lagged observations  $p$  as dimension of input vector can be seen as the most significant step of the model construction process because this parameter has a great effect on a determination of nonlinear auto-correlation pattern. Unfortunately, again there is no officially theoretical introduction for  $p$  selection but the practical experiments. All in all, a process of selecting a proper number of hidden nodes ( $q$ ) and dimension of input vector (lagged observations,  $p$ ) in practice would be determined as a main challenge because there is no principle-based guidance but experiences for choosing process of  $q$  and  $p$ . In terms of training a neural network, a backpropagation (BP) is probably a dominant approach which has been used to perform neural networks' model.

For neural network processing, the dataset is normally separated to three subset data that are used for three steps of training, test, and validation. In contrast, the process of specification, estimation, and validation of ARIMA model just often uses a whole dataset because the form of model is typically pre-identified before estimating data to get the model order. The backbone assumption for this process of designing ARIMA model is that that if the model is adapted to the time series data, it is capable to use for forecasting (Makridakis et al., 1982). Meanwhile, an estimation of the nonlinear form and the order of ANNs must emerge from the data which is overfitted by the model as a result. The major similarities between ARIMA model and ANNs are that they both consist of plentiful divisions of various models in which each has specific order, and the differentiating of data which is normally large enough should be made to gain best outcomes.

### Hybrid Model of ARIMA and ANNs

The hybrid model of ARIMA and ANNs has proven that it is a reliable measurement in the field of empirically statistics practice to explore the linear and nonlinear relationships lying in a certain dataset. Since ARIMA models is inadequate as estimating the correlation of nonlinear pattern, while ANNs model's outcomes for an approximation of linear pattern is insufficient reliable (Denton, 1995), (Markham & Rakes, 1998), so the hybrid model is likely to be a more appropriate solution to investigate both linear and nonlinear structures.

A well-known formation of hybrid model known is the additive model as follows:

$$\text{Additive Model : } y_t = L_t + N_t \quad (5)$$

where  $L_t$  and  $N_t$  are linear and nonlinear components respectively that must be explored from the dataset.

The performance of hybrid model includes (i) Applying ARIMA model to the linear part of time series as assumed as  $\{y_t; t = 1, 2, \dots\}$  to produce a series of forecast  $\{\widehat{L}_t\}$  which is used to compute the residuals  $\{e_t\}$  that enclose only a nonlinear part as follows based on the forms of additive model:

$$e_t = y_t - \widehat{L}_t \quad (6)$$

where  $e_t$  is the estimated residual at time  $t$  from the ARIMA (linear) model

The outcome of first stage is the nonlinear time series which would be used to integrate into ANNs in the second stage. In turn, ANNs performs the forecasts of nonlinear parts  $\{\widehat{N}_t\}$  with the inputs of  $\{e_t\}$  series as formed as follows:

$$e_t = f(e_{t-1}, e_{t-2}, \Lambda, e_{t-n}) + \varepsilon_t \quad (7)$$

This stage is known as an error correction of time series forecast in the ANNs with the ARIMA-based result. In forms of additive model, it has the equation helping to generate the forecast as follows:

$$\widehat{y}_t = \widehat{L}_t + \widehat{N}_t \quad (8)$$

In conclusion, the implement of the hybrid model consists of two stages that are an identification of ARIMA model with measuring the corresponding parameters to obtain the nonlinear part, and a prediction of ANNs based on the nonlinear part.

### 3 Model Design and Data Description

#### Hybrid Model

##### *Autoregressive Integrated Moving Average Model*

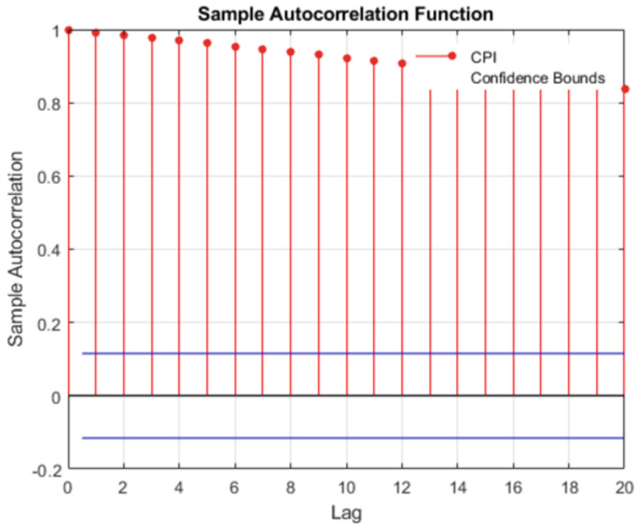
The data analysis of ARIMA model includes (i) Choosing the differentiation level  $d$  as to transform the data to the state of stationary; (ii) Evaluating the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to decide the most appropriate values of  $p$  and  $q$ . The data of Vietnam CPI from January 1995 to July 2022 (CPI series) provided by IMF data resource shows that there is seasonal pattern because the autocorrelation values at different lags are out of the confidence bounds as shown in the plot of ACF of CPI time series (Fig. 1). The result of first differentiation of CPI series (Fig. 2) shows a stationary pattern without the seasonal trend in the ACF and PACF, they both have large values at lags 1 (Figs. 3 and 4) within the 5% significance interval. Therefore, three possible models of ARIMA (0,1,1); ARIMA (1,1,0) and ARIMA (1,1,1) should probably be considered based on the criteria of AIC and BIC as shown in Table 1.

**Table 1.** Goodness of Fit of three possible ARIMA models

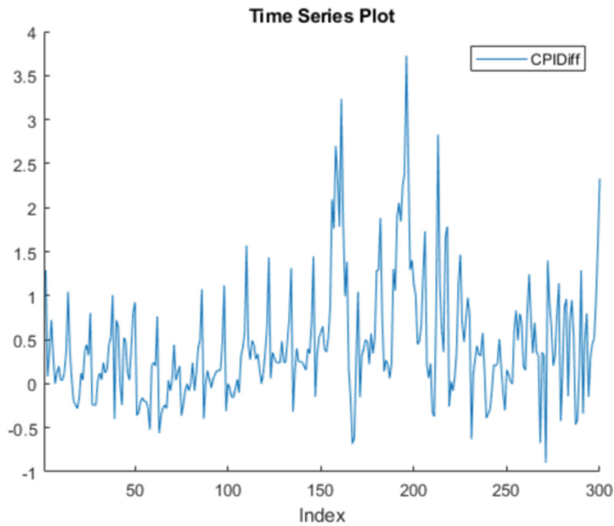
Model	AIC	BIC
ARIMA (0,1,1)	228.5305	237.8664
ARIMA (1,1,0)	233.7970	243.1149
ARIMA (1,1,1)	230.4307	242.8544



According to the figures of AIC and BIC, the ARIMA (0,1,1) is the best model in accordance with the data, since it has the smallest value of AIC and BIC. The distribution approach of ARIMA model applied in this research is Gaussian distribution.



**Fig. 1.** ACF of CPI time series



**Fig. 2.** The first differentiation of CPI series

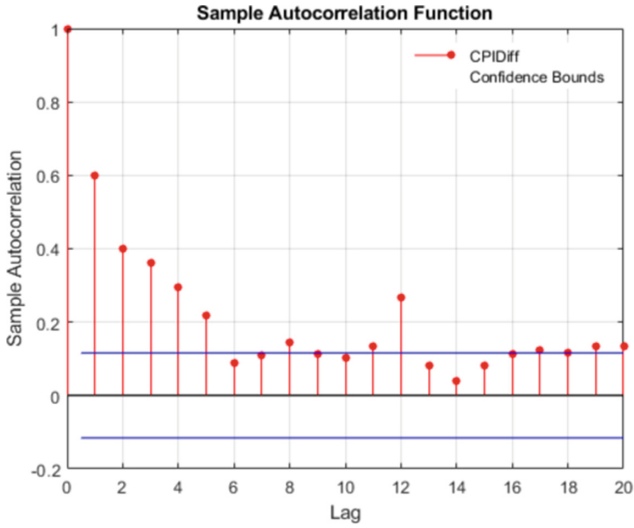


Fig. 3. ACF of first differentiation of CPI series

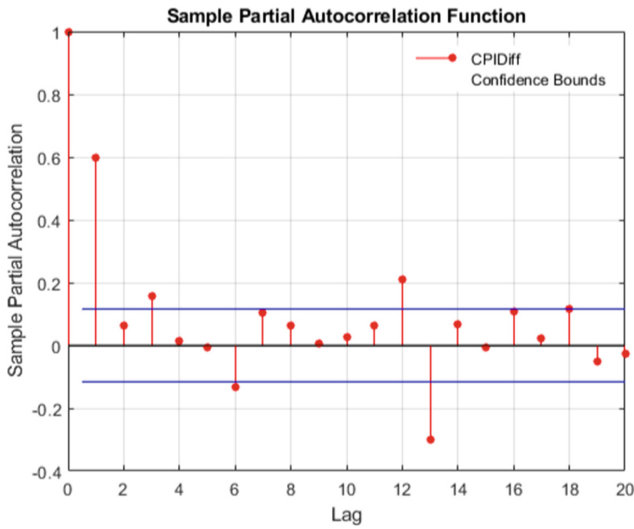


Fig. 4. PACF of first differentiation of CPI series

Table 1 shows a result of ADF test which aims to check a state of stationary of a time series data. The ADF test result proves that the first differentiation of Vietnam CPI series from 1995M01 to 2022M07 (Fig. 2) is the stationary (See Table 2).

**Table 2.** Augmented Dickey-Fuller Test

Test Statistic	Critical Value	Sig. level
<b>-6.6723</b>	-1.9417	0.0500

**Nonlinear Autoregressive Neural Network (NAR)**

Artificial neural networks (ANNs) which is developed based on the biological neurons system of human but in a form of mathematical model has a capability to verify nonlinear traits of time series data as modeling dynamic nonlinear time series. Basically, clusters of artificial neurons including several single neurons linked to each other in a network by weighted connections take responsibility to analyze a source of input which is activate by computed weights of mathematical function. Similarly, an output of the neural network is estimated by another activation function with a fitted threshold as follows:

$$y = f(b + \sum_i w_i x_i) \quad (9)$$

where  $y$  is the output;  $f$  is the activation function;  $b$  is the bias for the neuron algorithm which allows the signal to surpass the threshold of activation function;  $w_i$  are the weights;  $x_i$  are the inputs.

The Nonlinear Autoregressive Neural Network (NAR) has been developed based on a linear autoregressive model added the connections of feedback and widely used to predict multi-steps ahead of time series with past values of a real time series as the inputs by an equation as follows:

$$y_t = f(y_{t-1}, \dots, y_{t-d}) + \varepsilon_t \quad (10)$$

where  $f$  is the nonlinear function, in which the forecast values depend on series of past observations;  $y_{t-1}, y_{t-2}, y_{t-d}$  are feedback delays;  $d$  is the time delay.

The NAR is designed and trained in an open loop with target values as a response so as to ensure more appropriate quality approximate to the real number in training. Next, the network is transformed to a closed loop in which the predictions are used as new source of inputs. The process of training for the neural network is to estimate the model with using the optimization of the network weights and neuron bias as criteria.

The more detailed equation of NAR model as follows:

$$y_t = \alpha_0 + \sum_{j=1}^k \alpha_j \vartheta \left( \sum_{i=1}^a \beta_{ij} Y_{t-1} + \beta_{0j} \right) + \varepsilon_t \quad (11)$$

where  $k$  is the number of hidden layers of activation function  $\vartheta$ ;  $\alpha_j$  is the weighted link between the hidden unit  $j$  and the output unit  $\beta_{ij}$  is the parameter interacting with the weighted link between the input unit  $i$  and the hidden unit  $j$  with  $a$  entries;  $\beta_{0j}$  and  $\alpha_0$  are the constants to the hidden unit  $j$  and the output unit, respectively.

The dataset is separated into three categories based on the objective of performance as follows:

- Training: 70% of the dataset is used for the training process with an adjustment of the network based on its error.

- Validation: 15% of the dataset is served for the validation process including the network generalization and training stopping.
- Testing: 15% of the dataset is for the testing process as to optimize the network without affecting the training process, so that it could independently value the network performance during and after training.

### *The Selection of Hidden Layer Number*

This research applies the model with three layers of back-propagation network including input layer, output layer, and a single hidden layer in which a selection of the number of neurons majorly affects the structure of neural network because Cybenko (1989), Hornik et al. (1989), G. Zhang et al. (1998) demonstrated that a certain neural network model probably need only one single hidden-layer to efficiently investigate whatever complex non-linear relationship with having the accurate outcomes. While the Tan-Sigmoid function is a form of transformation from the hidden layer, the Sigmoid function is set for the output layer.

The structure of neural network is [2,10,2], in which the one-single hidden layer includes 10 neurons (as default layer size of NAR model in Matlab), the number of neurons of input and output layers have the same number of neurons of 2. The time delay (lags) is 2 as default setting of NAR model in Matlab.

### *The Evaluation Criteria*

For this research, the model performance criteria of neural network, with using the NAR model in Neural Network Toolbox of MatLab, are Mean-Squared Error (MSE with the maximum value of MSE is set at 0.001 (or 1000 iterations), the times of replication to get the averages of MSE, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are set at 10 for training process.

Regarding NAR model, this paper assesses an efficiency of the time-delay  $d$  in connection with the performance of the training process by using MSE function. The value of time-delay  $d$  is set at 2 as the default of NAR model. The applied function of training process is Bayesian Regulation backpropagation with the stopping point set at the finish of generalization improvement as the MSE of the validation dataset shows an increase.

### *Proposed Hybrid Model for Forecasting*

The CPI time series is nonstationary at the first hand as shown in Fig. 5, thus the hybrid model of ARIMA and NAR models is suggested to produce the forecast with a main objective is to process the nonstationary dataset by ARIMA model to gain the residuals series which is used as the input for NAR model. This process includes three stages:

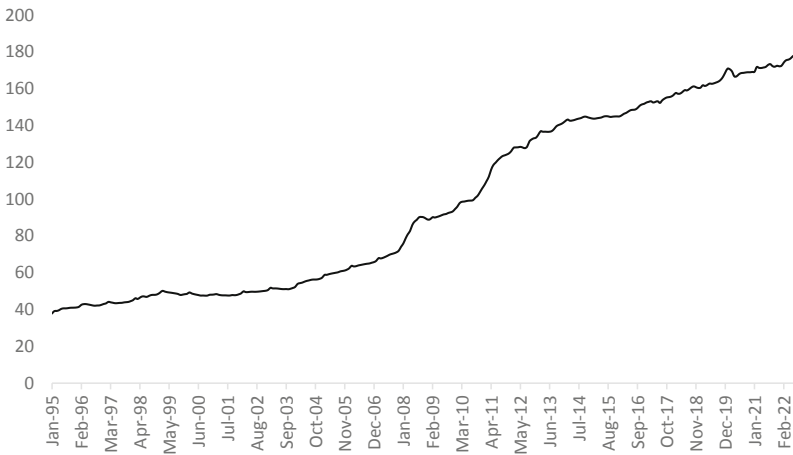
Stage 1: The residuals series (nonlinear part) produced by ARIMA (0,1,1) model is collected to use as the input of NAR model in Stage 2.

Stage 2: The NAR model is trained, validated, and tested with the input of the residual series gained in Stage 1.

Step 3: The fitted NAR model is applied to produce the multiple-step prediction of CPI series and compare with its actual values.

**Data Description**

Forecasting the fluctuation of CPI which reflects an inflation known as one of most significant factors of the economy has been still a tough challenge in the field of economical econometrics. For the theoretical perspective, although several linear and nonlinear models have been progressive, just some of them are likely to achieve an effectiveness in the outcomes of out-of-sample prediction in comparison with the simple-random-walk model. The data used in this paper is the monthly series of Vietnam CPI from January 1995 (1995M1) to July 2022 (2022M7) applied to the process of training, testing and validation of ARIMA, ANNs and the hybrid models. The illustration of the monthly observations of Vietnam CPI is illustrated in Fig. 1 showing a non-stationary pattern. Since the CPI series is non-stationary at first hand, thus the data is differentiated to become a stationary series to meet the basic assumption of ARIMA model. The dataset is separated into two sets that are a training sample including values from January 1995 to December 2008 (168 observations), and a testing sample including figures from January 2009 to July 2022 (163 observations) shown in Table 1. The function of training data set is for a process of model development and the test data set is for fitted model evaluation which shows the gap between the actual observations and forecasting values.



**Fig. 5.** Vietnam CPI monthly series (1995M01 – 2022M07)

**Table 3.** Data sample

Time period			
	Jan 1995 – Jul 2022	Jan 1995 – Dec 2008	Jan 2009 – Jul 2022
Series	Sample size	Training set	Test set
Vietnam CPI	319	168	163

## 4 Research Result

Tables 3 and 4 present the result of ARIMA model, and Goodness-of-fit of ARIMA and NAR models. Figure 6 shows the performance result of NAR model.

**Table 4.** Result of ARIMA (0,1,1)

Parameter	Value	Standard Error	t Statistic
<b>Constant</b>	- 0.0109	0.0237	- 0.4601
<b>MA{1}</b>	- 0.4177	0.0562	- 7.4300
<b>Variance</b>	0.2219	0.0192	11.5837

Regarding the NAR model, Table 5 shows the MSE which is used to evaluate the performance of model and regression R-value reflecting the correlation between outputs and responses of the training results and additional test results.

**Table 5.** Training results and Additional test results of NAR model

	Observation	MSE	R-value
<b>Model training</b>			
Training	113	0.3738	0.9997
Validation	24	0.8711	0.9988
Test	24	0.2210	0.9998
<b>Additional test</b>	163	34.2671	0.9835

**Table 6.** Goodness of Fit of ARIMA and NAR models

Model	AIC	BIC	Bayes factor
ARIMA	216.428	228.924	0.083*
NAR	128.476	73.699	

\*Bayes factor is computed based on the BIC of two models: ARIMA (0,1,1) and ARIMA (1,1,1)

The Bayes factor of ARIMA model shown in Table 6 is computed based on the equation of (Wagenmakers, 2007), the figure of 0.083 suggests that there is strong evidence in favor of the ARIMA (0,1,1) rather than others according to the interpretation of Bayes factor developed by (Jeffreys, 1961).

The forecast result of ARIMA (0,1,1), NAR and Hybrid models are illustrated in Fig. 7. The plots indicate that while neither ARIMA model nor NAR alone can produce the accurate prediction for the period of 163 months, the forecast of Hybrid model is greater accuracy, reflecting in the width of gap between the actual and the prediction lines. This probably proposes that either ARIMA model or NAR model has insufficiently statistical reliability to explore both the linear and the nonlinear parts in the time series data in order to produce highly accurate prediction. Meanwhile, the forecast of the Hybrid model reveals that as integrating two models together, the overall forecasting errors can be significantly minimized.

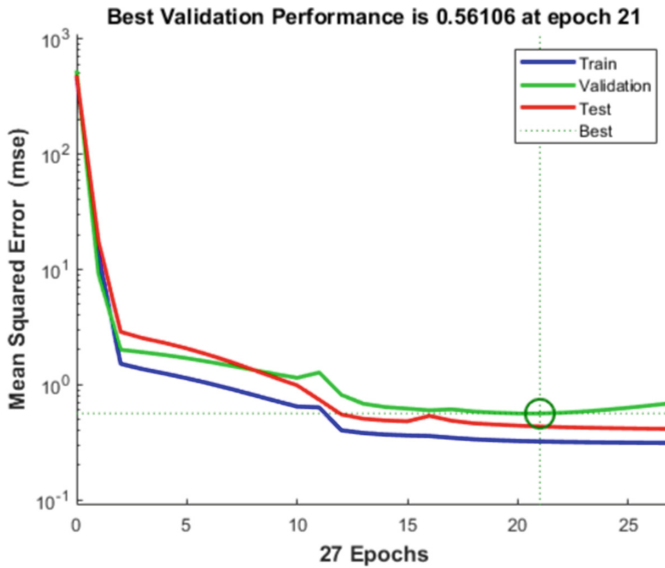


Fig. 6. Performance result of NAR model based on MSE

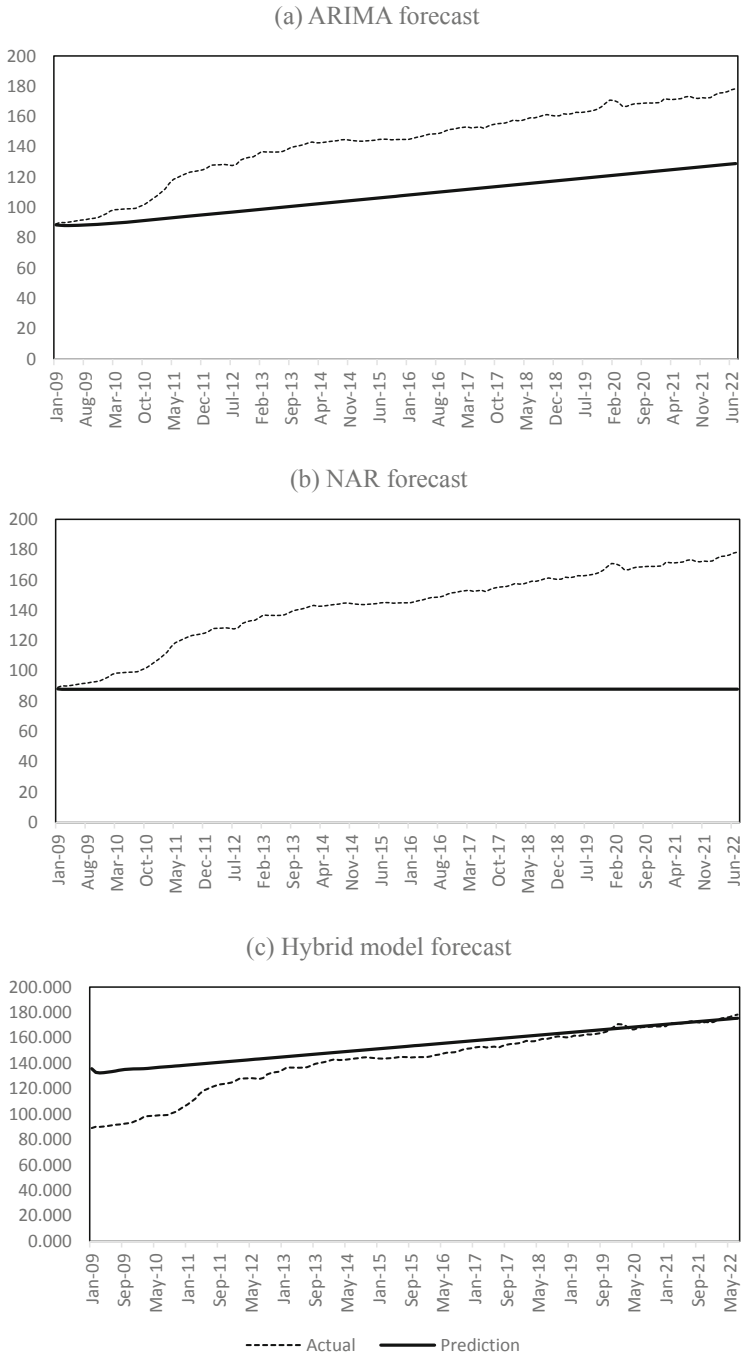


Fig. 7. (a) ARIMA (0,1,1) model forecast; (b) NAR model forecast; (c) Hybrid model forecast



## 5 Conclusion

While the ARIMA model has demonstrated its considerable advances in the field of time series forecasting, the artificial neural network model has also shown a great in that research area with its feature of exploring nonlinear pattern. However, none of them can be seen as the best model which can be used for every type of time series dataset. That's why this paper suggests a hybrid model which integrates ARIMA and ANNs models so as to produce more accurate prediction for any time series data. In other words, it takes full advantage of the capability of processing the linear pattern of ARIMA and the nonlinear pattern of ANNs. The empirical result of Vietnam CPI prediction conclusively demonstrates that the hybrid model is likely to outperform a single ARIMA model or ANNs as forecasting a multiple-step prediction of time series data.

## References

- Bates, J.M., Granger, C.W.J.: The combination of forecasts. *J. Oper. Res. Soc.* **20**(4), 451–468 (1969). <https://doi.org/10.1057/JORS.1969.103>
- Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford university press (1995)
- Box, G.: Box and Jenkins: time series analysis, forecasting and control. In: *A Very British Affair*, pp. 161–215. Palgrave Macmillan UK (2013). [https://doi.org/10.1057/9781137291264\\_6](https://doi.org/10.1057/9781137291264_6)
- Chatfield, C.: The future of the time-series forecasting. *Int. J. Forecast.* **4**(3), 411–419 (1988)
- Clemen, R.T.: Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.* **5**(4), 559–583 (1989)
- Cybenko, G.: Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2**(4), 303–314 (1989). <https://doi.org/10.1007/BF02551274>
- Denton, J.W.: How good are neural networks for causal forecasting? *J. Bus. Forecast.* **14**(2), 17 (1995)
- Ginzburg, I., Horn, D.: Combined neural networks for time series analysis. *Adv. Neural Inf. Proc. Syst.* **6**, 8 (1993)
- Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**(5), 359–366 (1989). [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- Jeffreys, H.: *Theory of probability* (3rd ed.), 432. Oxford university press, MR0187257 (1961)
- Jenkins, G.M.: Some practical aspects of forecasting in organizations. *J. Forecast.* **1**(1), 3–21 (1982)
- Kodogiannis, V., Lolis, A.: Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural Comput. Appl.* **11**(2), 90–102 (2002). <https://doi.org/10.1007/S005210200021>
- Kohzadi, N., Boyd, M.S., Kermanshahi, B., Kaastra, I.: A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing* **10**(2), 169–181 (1996). [https://doi.org/10.1016/0925-2312\(95\)00020-8](https://doi.org/10.1016/0925-2312(95)00020-8)
- Luxhøj, J.T., Riis, J.O., Stensballe, B.: A hybrid econometric-neural network modeling approach for sales forecasting. *Int. J. Prod. Econ.* **43**(2–3), 175–192 (1996). [https://doi.org/10.1016/0925-5273\(96\)00039-4](https://doi.org/10.1016/0925-5273(96)00039-4)
- Makridakis, S., et al.: The accuracy of extrapolation (time series) methods: results of a forecasting competition. *J. Forecast.* **1**(2), 111–153 (1982). <https://doi.org/10.1002/FOR.3980010202>
- Makridakis, S., et al.: The M2-competition: a real-time judgmentally based forecasting study. *Int. J. Forecast.* **9**(1), 5–22 (1993)

- Markham, I.S., Rakes, T.R.: The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Comput. Oper. Res.* **25**(4), 251–263 (1998)
- Naftaly, U., Ginzburg, I., Horn, D., Intrator, N.: Averaged and decorrelated neural networks as a time-series predictor. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440–5), 2, pp. 429–432 (1994)
- Palm, F.C., Zellner, A.: To combine or not to combine? Issues of combining forecasts. *J. Forecast.* **11**(8), 687–701 (1992). <https://doi.org/10.1002/FOR.3980110806>
- Trippi, R.R.: Neural networks finance and investment using artificial intelligence to improve real-world performance. In: Trippi, R.R., Turban, E.: (z-lib.org) (1996)
- Sfetsos, A., Siriopoulos, C.: Time series forecasting with a hybrid clustering scheme and pattern recognition. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **34**(3), 399–405 (2004). <https://doi.org/10.1109/TSMCA.2003.822270>
- Tsaih, R., Hsu, Y., Lai, C.C.: Forecasting S&P 500 stock index futures with a hybrid AI system. *Decis. Support Syst.* **23**(2), 161–174 (1998)
- Wagemakers, E.-J.: A practical solution to the pervasive problems of values. *Psychon. Bull. Rev.* **14**(5), 779–804 (2007)
- Walls, L.A., Bendell, A.: Time series methods in reliability. *Reliab. Eng.* **18**(4), 239–265 (1987). [https://doi.org/10.1016/0143-8174\(87\)90030-8](https://doi.org/10.1016/0143-8174(87)90030-8)
- Wang, J.-H., & Leu, J.-Y.: Stock market trend prediction using ARIMA-based neural networks. In: Proceedings of International Conference on Neural Networks (ICNN'96), 4, pp. 2160–2165 (1996)
- Wedding, D.K., Cios, K.J.: Time series forecasting by combining RBF networks, certainty factors, and the Box-Jenkins model. *Neurocomputing* **10**(2), 149–168 (1996). [https://doi.org/10.1016/0925-2312\(95\)00021-6](https://doi.org/10.1016/0925-2312(95)00021-6)
- Guoqiang Zhang, B., Patuwo, E., Hu, M.Y.: Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.* **14**(1), 35–62 (1998). [https://doi.org/10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)
- Zhang, G.P., Patuwo, B.E., Hu, M.Y.: A simulation study of artificial neural networks for nonlinear time-series forecasting. *Comput. Oper. Res.* **28**(4), 381–396 (2001)
- Zhang, P.G.: Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**, 159–175 (2003). [https://doi.org/10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0)



# A Bayesian Approach to Determinants of Capital Structure of Listed Construction Firms in Vietnam

Nguyen Duc Trung<sup>(✉)</sup>, Nguyen Ngoc Thach<sup>(✉)</sup>, and Bui Dan Thanh<sup>(✉)</sup>

HCMC University of Banking, Ho Chi Minh, Vietnam  
{ trungnd, thachnn, thanhbd }@buh.edu.vn

**Abstract.** This study aims to find out the factors affecting the capital structure of construction enterprises listed in Vietnam Stock Exchange in the period from 2014 to 2019. Using secondary data from 94 construction enterprises listed, including 472 observations with Bayesian regression technique to find out the factors affecting the overall debt ratio of enterprises. Regression results show that there are 6 important factors affecting capital structure which are enterprise size (SIZE), liquidity (LIQ), profitability (ROA), corporate income tax rate (TAX), age of the company (AGE) and growth opportunity (GROW). From there, business managers can refer to the research results to make decisions on capital structure, ensuring that it is consistent with the development goals of enterprises in the construction industry.

**Keywords:** Capital structure · construction business · Bayesian regression · Vietnam

## 1 Introduction

According to the 2011–2020 socio-economic development strategy report, the construction industry is an economic sector with an important strategic position and role in the construction and development of the country, always contributing over 5% GDP per year. According to statistics from the Research Institute for Infrastructure Development and Urban Planning, from mid-2018 until now, the growth rate of the construction and construction and infrastructure sectors has increased in 2018–2019, reaching 9.2–9.5%. In 2019, up to now, the construction market has followed a completely different trajectory, although the industry growth rate is only 7.02%. However, besides the achievements of Vietnam’s construction industry, enterprises have also faced many difficulties. The bad effects of the economic crisis, the Covid-19 pandemic, fluctuations in interest rates and inflation as well as the State’s management policies have directly affected the Construction industry. Therefore, studying the factors affecting the capital structure of construction enterprises is very necessary and will help enterprises maximize profits and limit financial risks. Therefore, the authors study the factors affecting the capital structure of construction companies listed on the Vietnamese stock market, which is an academic contribution to help enterprises in making capital management decisions.

## 2 Literature Review

### 2.1 The Traditional Theory of Capital Structure

David Durand (1952) is the first work on capital structure of enterprises with assumptions such as: Enterprises operate in an environment with corporate income tax, Financial markets are imperfect, and Enterprises have potential risk of financial distress due to the use of debt. When a business starts to take out debt, debt often costs less than equity. However, when the business increases the ratio of debt to equity, the level of risk also increases, forcing the owners to increase the rate of return. Thus, reducing the business value. According to this theory, there exists an optimal capital structure that maximizes firm value and helps to minimize WACC.

The main problem with the traditional view is that there is no theoretical basis for how much the cost of equity should increase as a result of the debt-equity ratio or by how much the cost of debt should increase due to risk. Therefore, M&M Theory was born on the basis of providing evidence as well as adding to the shortcomings that this view lacks (Brigham and Houston 2009).

### 2.2 The Modern Theory of Capital Structure (M&M Theory)

Contrary to the traditional view, two authors Modigliani and Miller (1958) have proposed a theory of the relationship between capital and firm value. To find out if the value of capital increases or decreases as the business increases or decreases borrowing. The theory is made with the following assumptions: No taxes (corporate income tax and personal income tax), No costs: transactions, financial distress and bankruptcy, all investors both individual and corporate investments have the same interest rates, and Capital Markets are perfect markets. M&M theory is stated into two important propositions: Statement (I)-Value of the enterprise and Proposition (II)-Cost of capital. The two propositions are considered in two cases, respectively: with tax and without tax.

### 2.3 Trade-off Theory of Capital Structure

Based on the M&M theory, the trade-off theory considers the impact of taxes and the cost of financial distress. Initiated by Kraus and Litzenberger (1973) and developed by Myers (1977) the firm should only use a certain amount of debt to maximize firm value, in contrast to the value M&M theory. The higher the company, the more it is used. The trade-off theory has shown that the target capital structure is the point at which the benefits from the tax shield can offset the costs of financial distress. However, when the debt ratio rises to a certain level, the cost of financial distress will outweigh the benefit of the tax shield from interest. From there, the company value will decrease and increase the probability of bankruptcy.

$$\begin{aligned} \text{Value of the debt firm} &= \text{The value of the unlevered firm} + \text{Present value of the tax shield} \\ &\quad - \text{Present value of the cost of financial distress} \end{aligned}$$

Factors affecting capital structure from a structural trade-off point of view include: corporate income tax, financial distress costs, tangible fixed assets, company size and profitability.

The trade-off theory has explained the dead side of the M&M theory about the cost of financial distress of debt-ridden firms. However, there are also many things that the trade-off theory cannot explain, such as why some businesses are still successful, good business results when borrowing very little debt; or in fact, when the company's stock price is high and the firm is in need of external financing, the company is more likely to issue shares (rather than take out debt).

## 2.4 Pecking Order Theory

Besides the above theories, the pecking order theory developed by Stewart Myers and Nicolas Majluf (1984) goes in another direction when it says that there is no optimal capital structure, but only an order when using investments. The study divides funding sources into: internal capital (contributed capital and retained earnings) and external capital (borrowed capital and new share issuance).

According to Myers and Majluf, based on the information asymmetry between financial managers and outside investors. Managers will have more information than investors, so investors will often demand higher discounts, making the cost of raising outside capital will be higher. This leads to the formation of a funding priority order.

Although the pecking order theory explains some aspects that affect the decision to choose financing sources of enterprises, this theory still has many limitations when it does not explain the impact of taxes, bankruptcy cost, the cost of issuing securities to the enterprise's debt.

## 2.5 The Market Timing Theory

According to market timing theory, capital structure is the cumulative result of past efforts to time the stock market. That is, there is no optimal capital structure according to the Trade Off Theory. Research by Baker and Wurgler (2002) shows a new perspective on the problem of capital structure, when it is said that the value of enterprises depends on two factors: stock price and time to enter the market.

Thus, there are many theories of capital structure that have been presented and applied. This study focuses on applying two theories of capital structure, namely the trade-off theory of capital structure and the pecking order theory.

## 2.6 Comprehensive Study

*Bhaduri* (2002) argued that non-debt tax shields are good substitutes for tax benefits from debt, so firms with large non-debt tax shield benefits will borrow less. The study also suggests that large-scale enterprises will tend to diversify their mobilized capital sources, so they are less prone to financial crises, in other words, there is a positive relationship of enterprise size to capital structure.

**Chen (2004)** carried out on 77 large companies whose shares are listed on the Shanghai Stock Exchange, China. The author has based on the trade-off theory and pecking order theory to determine the factors affecting the capital structure of listed companies, including: profitability, growth ability, tangible fixed assets, financial distress costs and tax shields on the capital structure of firms. Chen's research results show that profitability and firm size have a negative impact on capital structure, while growth rate and tangible fixed assets have a positive impact on debt ratio. At the same time, according to Chen, pecking order theory explains the research results better than trade-off theory.

**Wanrapee Banchuenvijit (2009)** carried out on 81 companies listed on the Stock Exchange of Thailand from 2004 to 2008. With 5 factors included in the model, including profitability, firm size, the ratio of tangible fixed assets, the growth rate of assets, the volatility of operating profit. The results show that there are three factors that are statistically significant at the 1% level, namely: profitability, fixed assets have a negative relationship with the debt ratio, and firm size is positively related. in the same direction as the debt ratio.

**Tran Dinh Khoi Nguyen and Ramachandran (2006)** carried out, to test the factors that play an important role in the capital structure decision of 558 small and medium enterprises in the period 1998–2001. The results show that the variables of enterprise size, business risk, relationship with banks and growth rate of revenue are positively correlated with capital structure of enterprises. In contrast, profitability and asset structure have a negative effect on the debt to total assets ratio of the firm.

**Dang Thi Quynh Anh and Quach Thi Hai Yen (2014)** examined the impact of 10 factors affecting the capital structure of enterprises listed on the Ho Chi Minh Stock Exchange in the period 2010–2013. The research results show that the three factors that have the strongest influence on the capital structure of enterprises are the profitability ratio, the size of the enterprise and the corporate income tax rate. Profitability and tax rates have a negative impact on the financial leverage of enterprises, while the size of enterprises has a positive effect.

**Nguyen Thi Nhu Quynh, Le Dinh Luan, Le Hoang Vinh (2020)** analyze the factors affecting the capital structure of 148 non-financial enterprises listed on HOSE through short-term financial leverage and ratio long-term financial leverage in the period from 2011–2018. Research has shown some interesting points, in the short term, the leverage ratio is affected by the factors of profit size, asset structure and corporate liquidity. In the long run, leverage is influenced by size, profitability, asset structure, growth opportunities, and liquidity. Taxes do not affect capital structure.

It is noteworthy that the aforementioned research used frequency approaches or descriptive analyses with suitably large sample sizes to analyze capital structure in sample enterprises. Based on a dataset of 94 construction companies listed on the Vietnam Stock Exchange between 2014 and 2019, this study used Bayesian logistic regression with informative priors. The research has made the following contributions, as expected: (i) Business managers can use the research findings to inform their capital structure decisions, ensuring that they align with the objectives of businesses in the construction sector for growth; (ii) By using Bayesian MCMC simulations in informative (thoughtful) prior settings, our findings enable a generalized conclusion that, in contrast to frequentist approaches, Bayesian estimation using thoughtful priors can provide meaningful results.

### 3 Model and Data

#### 3.1 General Model

$$Y_{it} = \beta_0 + \sum \beta_i X_{it} + u_{it}$$

In which:  $i$ : the  $i$ -th cross unit and  $t$  is the  $t$ -th time;  $Y_{it}$  is the dependent variable;  $X_{it}$  is the independent variable;  $\alpha$ : coefficient of freedom,  $\beta$ : coefficient of regression,  $u_{it}$ : residual.

Based on empirical studies in the world and in Vietnam, the authors find that the number of variables as well as the way to measure the variable and the resulting direction of the impact of the variables (factors) on capital structure is varied across studies. However, these studies all selected a number of factors affecting the capital structure of enterprises such as firm size, asset structure, liquidity, profitability ratio, tax growth opportunities and business age. These variables all have the ability to collect data and all have economic significance, are correlated and explain the research problem. Therefore, the author has built a research model and introduced variables that affect the capital structure of listed construction companies in Vietnam on the basis of selecting the impact variable of previous empirical studies. The research model of the topic is as follows:

$$\begin{aligned} \text{TLEV}_{it} = & \mathbf{a} + \mathbf{b}_1 \text{SIZE}_{it} + \mathbf{b}_2 \text{TANG}_{it} + \mathbf{b}_3 \text{LIQ}_{it} + \mathbf{b}_4 \text{ROA}_{it} + \mathbf{b}_5 \text{GROW}_{it} \\ & + \mathbf{b}_6 \text{TAX}_{it} + \mathbf{b}_7 \text{AGE}_{it} + \mathbf{u}_{it} \end{aligned}$$

In which:

$\text{TLEV}_{it}$ : overall debt ratio of enterprise  $i$  at year  $t$  (Total debt to total assets)

$\text{SIZE}_{it}$ : Size of enterprise  $i$  at year  $t$  (Natural Logarithm of Total Assets)

$\text{TANG}_{it}$ : Asset structure of enterprise  $i$  at year  $t$  (The ratio of tangible fixed assets to total assets)

$\text{LIQ}_{it}$ : Liquidity of enterprise  $i$  at year  $t$  (Ratio of current assets to current liabilities)

$\text{ROA}_{it}$ : Profitability of enterprise  $i$  in year  $t$  (Rate of profit after tax to total assets)

$\text{GROW}_{it}$ : Growth rate of enterprise  $i$  in year  $t$  (Difference of total revenue at the end of the period and total revenue at the beginning of the period over total revenue at the beginning of the period)

$\text{TAX}_{it}$ : Actual tax rate of enterprise  $i$  in year  $t$  (Corporate income tax rate on pre – tax profit of that enterprise)

$\text{AGE}_{it}$ : Age of enterprise  $i$  at year  $t$  (Logarithm of year of study minus year of establishment)

#### 3.2 Variables and Hypotheses

##### Dependent Variable (TLEV)

In this study, capital structure is determined by the total debt ratio (TLEV).

$$\text{TLEV} = (\text{Total Liabilities})/(\text{Total Assets})$$

##### Independent Variables

Within the scope of the study, the independent variables of the model only focus on

the factors from the internal resources of the enterprise affecting the employees, not considering the macro factors.

**Enterprise Size (SIZE):** Enterprise size is measured by the value of total assets of the enterprise. However, because the value of total assets is large, the topic converts the natural logarithm of total assets to reduce the value difference between the variables.

$$\text{SIZE} = \text{Log}(\text{Total Assets})$$

As stated in the theoretical basis, the size of the firm to the shareholder can be a positive relationship (according to the trade-off theory) or a negative relationship with the debt coefficient (according to the pecking order theory). But most recent studies such as Dang Thi Quynh Anh and Quach Thi Hai Yen (2014) show a positive relationship between enterprise size and debt ratio. *The first hypothesis (H<sub>1</sub>): Firm size has a positive relationship with debt ratio.*

**Structure of Tangible Assets (TANG):** Tangible fixed asset is a variable reflecting the structure of assets of a construction enterprise, measured by the ratio between tangible fixed assets and total assets of the enterprise.

$$\text{TANG} = (\text{Tangible fixed assets})/(\text{Total assets})$$

Tangible fixed assets characterize the willingness of enterprises to mortgage before loans. Collateral is a good, important condition for creditors to consider credit decisions. According to the trade-off theory and the results of studies by Chen (2004), Nguyen Thi Nhu Quynh et al. (2020), the ratio of tangible fixed assets (TANG) has a positive relationship with the financial leverage of enterprises. In this study, the author also predicts that tangible fixed assets have a positive relationship with the debt ratio of enterprises. *The second hypothesis (H<sub>2</sub>): The structure of tangible fixed assets has a positive relationship with the debt ratio.*

**Liquidity Variable (LIQ):** The liquidity of assets of construction enterprises is measured by the ratio between short-term assets and short-term liabilities of enterprises.

$$\text{LIQ} = (\text{Current Assets})/(\text{Current Liabilities})$$

According to pecking order theory, the liquidity of enterprises is negatively related to the debt ratio. Because businesses have abundant liquidity, they can use these assets to finance their investments as retained earnings, without the need to raise external capital. *The third hypothesis (H<sub>3</sub>): Liquidity has a negative relationship with debt ratio.*

**Return on Assets (ROA):** The return on assets (ROA) ratio measures a company's ability to earn per dollar of assets. According to Nguyen Minh Kieu (2009), the formula for determining this ratio is by dividing net profit after tax by total asset value.

$$\text{ROA} = (\text{Profit after tax})/(\text{Total assets})$$



Profitability has both a positive effect (according to the trade-off theory) and a negative effect (according to the pecking order theory) on the use of debt of enterprises. According to most recent empirical studies, profitability and coefficient have an inverse relationship such as Wanrapee Banchuenvijit (2009), Tran Dinh Khoi Nguyen and Ramachandran (2006), Dang Thi Quynh Anh and Quach Thi Hai Yen (2014). The author also predicts that ROA has a negative relationship with the debt ratio of enterprises. ***The fourth hypothesis (H<sub>4</sub>): Profitability has a negative relationship with debt ratio.***

**Growth Rate Variable (GROW):** The Growth Rate Variable (GROW) is used to measure the effect of annual revenue growth on firm value. The value of the variable is determined:

$$\text{GROW} = (\text{Revenue}_n - \text{Revenue}_{(n-1)}) / (\text{Revenue}_{(n-1)})$$

Revenue growth is one of the top concerns of corporate managers, according to previous empirical studies, which have shown a positive relationship between growth rate and debt ratio like Chen (2004), Nguyen Thi Nhu Quynh, Le Dinh Luan and Le Hoang Vinh (2020). ***The fifth hypothesis (H<sub>5</sub>): Growth rate has a positive relationship with the debt ratio.***

**Corporate Income Tax Rate (TAX):** The corporate income tax rate is measured by the ratio between the payable corporate income tax on the pre-tax profit of the enterprise.

$$\text{TAX} = (\text{Corporate income tax payable}) / (\text{Profit before tax})$$

According to M&M theory, and trade-off theory, there is a positive relationship between tax and debt ratio as studied by Chen (2004). However, in recent years, studies have shown that there is an inverse relationship between corporate income tax rates and the use of debt by enterprises such as Dang Thi Quynh Anh and Quach Thi Hai Yen (2014), Le Quynh Anh and Quach Thi Hai Yen (2014), Thi Minh Nguyen (2016). Therefore, the author expects in this study the relationship between corporate income tax rate and debt ratio is negative. ***The sixth hypothesis (H<sub>6</sub>): The corporate income tax rate has a negative relationship with the debt coefficient.***

**Enterprise Age Variable (AGE):** The age of the company is determined by logarithm of the number of years from inception to the year of the study data collection. And the author predicts that in this study, the age of enterprises has a positive effect on the debt coefficient.

$$\text{AGE} = \text{Log}(\text{Year of Research} - \text{Year of Establishment})$$

***The Seventh Hypothesis (H<sub>7</sub>): Firm Age Has a Positive Relationship with Debt Coefficient (Table 1).***

### 3.3 Data

**Table 1.** Description of the model's variables, measurement methods and hypotheses

Variable	Description	Measurement	Hypotheses
<b>Dependent variable</b>			
<b>TLEV</b>	The total debt ratio	Total Liabilities/Total Assets	
<b>Independent variables</b>			
<b>SIZE</b>	Enterprise Size	Log(Total Assets)	H <sub>1</sub> : +
<b>TANG</b>	Structure of tangible assets	Tangible fixed assets/Total assets	H <sub>2</sub> : +
<b>LIQ</b>	Liquidity	Current Assets/Current Liabilities	H <sub>3</sub> : –
<b>ROA</b>	Return on Assets	Profit after tax/Total assets	H <sub>4</sub> : –
<b>GROW</b>	Growth Rate	(Next year's revenue – Previous year's revenue)/Previous year's revenue	H <sub>5</sub> : –
<b>TAX</b>	Corporate income tax rate	(Corporate income tax payable)/(Profit before tax)	H <sub>6</sub> : –
<b>AGE</b>	Enterprise age	Log (Year of Research – Year of Establishment)	H <sub>7</sub> :+

Source: Compiled by the author

### 3.4 Model Estimation Method

To evaluate the impact of foreign ownership on liquidity risk, the authors will make model estimation according to Bayesian approach. To conduct a Bayesian analysis, a priori information is required for the research model, but since most of the prior research was performed using a frequency approach, a priori information is not available. However, the research data of 472 observations is quite large, so the a priori information does not have a great influence on the posterior distribution. In this case, Block et al. (2011) proposed a standard Gaussian distribution with different a priori information (simulation of a priori information) and carried out Bayesian factor analysis to choose a simulation with the best priori news.

The simulations in Table 2 show decreasing levels of a priori information with Simulation 1 having the strongest a priori information and Simulation 5 having the weakest a priori information.

**Table 2.** Simulation of a priori information

Rational function	TLEV $\sim N(\mu, \sigma)$
A priori distribution	
Simulation 1	$\alpha \sim N(0, 1)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 2	$\alpha \sim N(0, 10)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha \sim N(0, 100)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 4	$\alpha \sim N(0, 1000)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 5	$\alpha \sim N(0, 10000)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$

Source: Compiled by the author

In the next step, the research team carried out Bayesian regression for the above simulations, then performed Bayesian factor analysis (Bayes Factors) and Bayes test model (bayestest model). These are the techniques proposed by StataCorp LLC (2019) to select the simulation with the best a priori information. Basically, the Bayesian factor will provide a tool to compare the probability of a particular hypothesis (a priori information) to the probability of another hypothesis. It can be understood as a measure of the strength of evidence in favor of a theory among competing (information a priori) theories. Accordingly, Bayesian analysis will provide average Log BF (Bayes Factor), Log ML (Marginal Likelihood) and average DIC (Deviance Information Criterion-information bias); The posterior Bayesian test will help compare the posterior probability of the simulations with different a priori information, accordingly, based on the research data combined with the proposed a priori information, we will choose the simulation has the greatest posterior probability  $P(M|y)$ .

In summary, in this study, the research team will build 5 simulations with 5 different a priori information, and Bayesian factor analysis and posterior Bayesian test will help to choose a simulation with suitable a priori information. The simulation selected will be the one with the largest Log BF, Log ML average, minimum DIC mean and the largest  $P(M|y)$ .

## 4 Research Results and Discussion

### 4.1 Results

**Table 3.** Bayes Factor analysis results

	Chains	Avg. DIC	Avg. log (ML)	Log (BF)	P (Mly)
Simulation 1	3	-597.617	265.344		0.9997
Simulation 2	3	-597.643	257.062	-8.282	0.0003
Simulation 3	3	-597.709	247.965	-17.379	0.0000
Simulation 4	3	-597.643	238.796	-26.548	0.0000
Simulation 5	3	-597.595	229.604	-35.740	0.0000

Source: Calculations of the author

Table 3 shows that simulation 1 meets the criteria to be the most suitable a priori information simulation. Moreover, the results of post-test also show that simulation 1 has superiority over other simulations, so simulation 1 with a priori information  $N(0,1)$  will be selected.

Bayesian analysis is simulated through the Markov chain Monte Carlo (MCMC), therefore, to ensure the stability of the Bayesian regression, the MCMC series must converge, which means that the MCMC series must ensure stationarity StataCorp LLC (2019) proposes that the MCMC series convergence test can be conducted through the convergence diagnostic graph.

According to StataCorp LLC (2019), the MCMC series convergence diagnostic graph includes trace plot, histogram, autocorrelation, and density estimation. The trace plot helps to track the historical display of a parameter value over the iterations of the series, Fig. 1 shows the trace plot fluctuates around the mean value, so the MCMC series is stationary, that is, reaching convergence conditions. Besides, the autocorrelation chart in the graphs only fluctuates around the level below 0.02, according to StataCorp LLC (2019) the autocorrelation chart fluctuates around the level below 0.02, showing the agreement with the density simulate the distribution and reflect all delays that are within the effective limit. According to StataCorp LLC (2019), the posterior distribution plot and density estimate show that the simulation of the shape of the normal distribution of the parameters, the histogram shape is uniform, it can be concluded that Bayesian regression ensure stability. Thus, the results from Fig. 1 show that the MCMC series meets the convergence condition.

In addition to graphical convergence diagnostics, StataCorp LLC (2019) also recommends testing through Mean Acceptance Rate; Average minimum efficiency; and Gelman-Rubin  $R_c$  max. Table 4 shows that the model's acceptance rate reaches 1, the model's minimum efficiency is 0.99, far exceeding the allowable level of 0.01. In addition, the maximum  $R_c$  value of the coefficients is 1, Gelman and Rubin (1992) argue that the diagnostic value  $R_c$  of any coefficient of the model greater than 1.2 will be

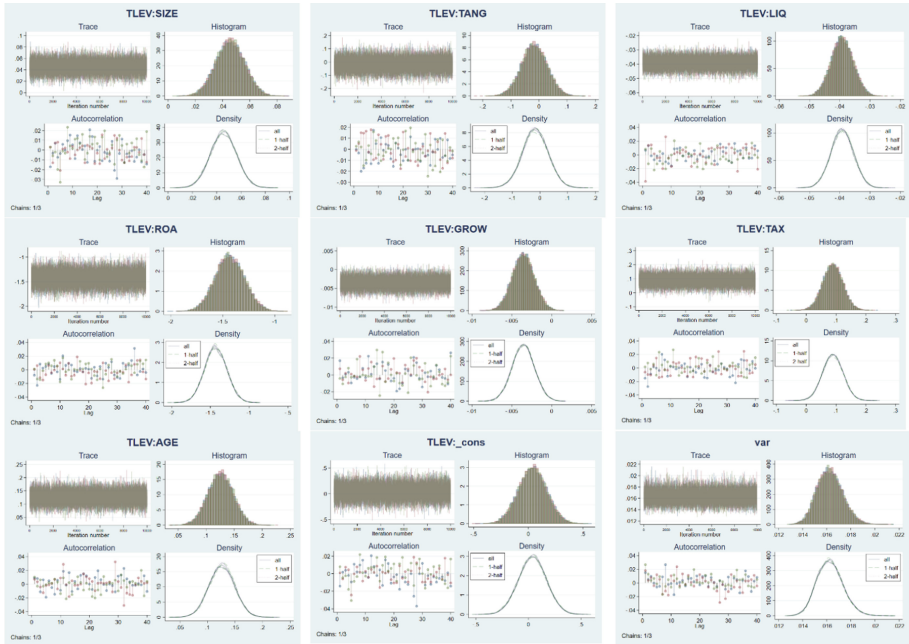


Fig. 1. Convergence diagnostic graph Source: Calculations of the author

Table 4. Regression results

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
SIZE	0.046	0.011	0.000	0.046	0.024	0.067
TANG	-0.017	0.047	0.000	-0.017	-0.108	0.073
LIQ	-0.039	0.004	0.000	-0.039	-0.047	-0.032
ROA	-1.432	0.142	0.001	-1.432	-1.712	-1.153
GROW	-0.004	0.001	0.000	-0.004	-0.006	-0.001
TAX	0.089	0.034	0.000	0.089	0.022	0.155
AGE	0.126	0.023	0.000	0.126	0.081	0.171
_cons	0.039	0.132	0.001	0.037	-0.218	0.300
var	0.016	0.001	0.000	0.016	0.014	0.019
Avg acceptance rate	1					
Avg efficiency min	0.9919					
Max Gelman-Rubin Rc	1					

Source: Calculations of the author

considered non-convergent. Thus, the values in Table 4 show that the MCMC series of the model satisfy the convergence requirements.

The regression results in the Table 4 have determined that the variables SIZE, TAX, AGE have a positive impact on the capital structure of construction firms while the variables TANG, LIQ, ROA, GROW have a negative impact on the capital structure of the construction industry. Besides determining the sign of the regression coefficients, unlike the frequency method, the Bayesian approach also allows us to calculate the probability of the occurrence of these effects (Table 5).

**Table 5.** Probabilistic test

	Mean	Std. Dev.	MCSE
{TLEV:SIZE} >0	1.000	0.000	0.000
{TLEV:TANG} <0	0.640	0.480	0.003
{TLEV:LIQ} <0	1.000	0.000	0.000
{TLEV:ROA} <0	1.000	0.000	0.000
{TLEV:GROW} <0	0.995	0.072	0.000
{TLEV:TAX} >0	0.996	0.065	0.000
{TLEV:AGE} >0	1.000	0.000	0.000

Source: Calculations of the author

## 5 Discussion

**Enterprise Size:** The positive effect of firm size on debt ratio, consistent with the capital structure trade-off theory and in accordance with the hypothesis  $H_1$  posed. The larger the business size, the stronger the financial potential, the lower the bankruptcy risk. In addition, large-scale enterprises have a better reputation in the debt market, gain trust from creditors, so they can easily access loans and reduce transaction costs when issuing long-term debt. Huang and Song (2001), Bhaduri (2002), Wanrapee Banchuenvijit (2009), Tran Dinh Nguyen Khoi and Ramachandran (2006), Harc (2015), Dang Thi Quynh Anh and Quach Thi Hai Yen (2014) gave the results similar to the research.

**Structure of Tangible Fixed Assets:** The results show that the level of impact of TANG is not really clear when its impact probability is only 64%, this is not true with the author's initial expectation, but to explain this, the construction industry in Vietnam is a project-based business, so it doesn't need a lot of fixed assets as collateral. It can be seen that in the capital structure of construction enterprises, mainly short-term loans, because the project-based business only borrows and calls for short-term investment capital, it does not need many fixed assets.

**Liquidity:** Liquidity has a negative relationship with the overall debt ratio of enterprises. This result is completely consistent with the original hypothesis and the point of view

of pecking order theory about shareholders, that enterprises with good liquidity will be more likely to convert short-term assets into cash to finance arising capital needs rather than borrowing from outside. Some studies by Le Thi Minh Nguyen (2016), Nguyen Thi Thuy Hanh (2018) also have results on the negative correlation between liquidity and corporate shareholders.

**Profitability:** ROA has the opposite and statistically significant effect on capital structure, high-profit enterprises in the construction industry listed in Vietnam use a lot of equity and little debt. Therefore, a firm with high profits will avoid taking on a lot of debt, this result is consistent with pecking order theory. That once again confirms, businesses with high profitability tend to finance with internal capital rather than external capital. Similar results are shown in studies Huang and Song (2001), Chen (2004), Wanrapee Banchuenvijit (2009), Tran Dinh Khoi Nguyen and Ramachandran (2006), Dang Thi Quynh Anh and Quach Thi Hai Yen (2014), Nguyen Thi Nhu Quynh, Le Dinh Luan and Le Hoang Vinh (2020).

**Growth Opportunity:** Growth opportunities have a negative impact on the capital structure of construction enterprises, although not in line with the author's expectations, it can be explained that construction enterprises with growth opportunities tend to use less debt to operate. It is also reasonable that when construction enterprises mainly use short-term loans, so there is great pressure on debt repayment, so when there is a good growth rate, construction enterprises will limit the use of loans, reduce financial costs for businesses. Research results agree with the opinion of Huang and Song (2001).

**Corporate Income Tax Rate:** The value of corporate tax rate measured by corporate income tax on EBT has a positive correlation with capital structure, although it is not consistent with the author's initial expectation, but it is consistent with the business situation of the company. Construction enterprises use a lot of debt. It also means that the higher the corporate tax rate, the higher the financial leverage of the business and vice versa. In addition, interest has created a "tax shield" for businesses, thereby creating higher business efficiency than using only equity. This assertion is also consistent with the results of Chen (2004).

**Age of Business:** The age of enterprises has a positive correlation with the capital structure of enterprises with a long history of operation in the market, the position of the enterprise has also been confirmed, the higher the prestige, the ability to borrow capital from the higher the regulations. Similar results were shown in the study of Mutalib (2011).

## 6 Conclusion and Policy Implications

The ultimate goal of corporate financial management is to maximize the value of the business, thereby maximizing the value for the owners of the business, this is done through minimizing the average cost of capital (WACC), including the cost of equity and the cost of debt. The research topic has identified the factors affecting the use of debt of construction enterprises, from which the managers can increase or decrease the level of debt use through affecting the factors of the model. In the current context of Vietnam's financial market, the authors propose some recommendations as follows:

**Adjust the Size of the Business Appropriately:** Enterprise size can bring positive effects, but at the same time, it can also become a burden of bankruptcy risk if enterprises do not have reasonable adjustment solutions. Enterprises should expand when there are many investment projects and vice versa, businesses with low debt ratio can scale synchronously to access more and more loan sources. However, when the debt ratio of enterprises increases, surpasses the alarming threshold in the context of the economy showing signs of decline or enterprises are facing difficulties, business managers should have solutions to adjust the size of the enterprise in order to adjust the target capital structure of the enterprise, avoiding the risk of bankruptcy.

**Improve Corporate Financial Management Capacity:** Construction enterprises are at the end of the growth period and entering the restructuring phase in the years 2014–2019, so financial management and maximizing corporate value of the company are very important. Therefore, leaders must really be aware of the role of financial management, as well as in-depth knowledge of the field of financial management to consider options for mobilizing and using funding for projects. Projects in an appropriate and effective manner and to limit financial risks. To do this, businesses need to specialize by separating the financial and accounting functions; at the same time consider using financial hedging tools such as financial derivatives.

**Improve the Efficiency of Production and Business Activities:** The profitability factor (ROA) has a negative impact on the target capital structure of the enterprise. Increased profitability will reduce the debt ratio of businesses. When businesses have abundant internal capital and have increasing profits, businesses will have the necessary financial autonomy. To achieve the above goals, businesses need to improve business efficiency and develop specific financial plans to avoid wasting capital as well as better manage and control production costs.

**Increase Equity Capital and Exploit more Capital Mobilization Channels:** The solution to increase equity capital will help improve the financial autonomy of businesses, which will help these businesses overcome difficult times when banks reduce lending limits, increase interest rates. Construction enterprises can increase their equity capital by ways such as increasing retained earnings, concentrating on collecting outstanding debts from projects, expanding scale, calling for more members as well as shareholders contribute capital, this will help the business have a huge amount of additional capital.

**Increase Transparency of Information:** One of the barriers that reduces the ability of enterprises to access loans is the problem of information disparity. This also directly affects the ability to win contracts of construction enterprises in particular and the access to investment projects of enterprises in general. Therefore, agencies and sectors need to continue to develop and improve regulations on publicity and transparency of information systems on both supply and demand of the market, in order to reduce the harmful effects of asymmetric information. The authorities also need to strengthen the collection and disclosure of information, and at the same time should build a national information infrastructure to help banks quantify the capacity and risks of businesses to adopt appropriate policies.



## References

- Anh, D.T., Yen, Q.T.: Factors affecting capital structure of enterprises listed on Ho Chi Minh stock exchange. *J. Dev. Integr.* (18) (2014)
- Baker, M., Wurgler, J.: Market timing and capital structure. *J. Financ.* **57**(1), 1–32 (2002)
- Bhaduri, S.N.: Determinants of corporate borrowing: some evidence from the Indian corporate structure. *J. Econ. Financ.* **26**(2), 200–215 (2002)
- Brigham, E.F., Houston, J.F.: *Fundamentals of Financial Management, Concise Edition*, 6th edn. Cengage Learning (2009)
- Chen, J.J.: Determinants of capital structure of Chinese-listed companies. *J. Bus. Res.* **57**(12), 1341–1351 (2004)
- Durand, D.: Costs of debt and equity funds for business: trends and problems of measurement. In: *Conference on Research in Business Finance*, pp. 215–262. NBER (1952)
- Field, A.: *Discovering Statistics Using SPSS for Windows: Advanced Techniques For Beginners*. Sage Publication, Great Britain (2000)
- Huang, G., Song, F.M.: The financial and operating performance of China's newly listed H-firms. *Pac. Basin Financ. J.* **13**(1), 53–80 (2005)
- Kieu, N.: *Basic corporate finance: theory and practice of applied management For Vietnamese enterprises*. Hanoi: Statistics, 884 p. (2009)
- Kraus, A., Litzenberger, R.H.: A state-preference model of optimal financial leverage. *J. Financ.* **28**(4), 911–922 (1973)
- Modigliani, F., Miller, M.H.: Corporate income taxes and the cost of capital: a correction. *Am. Econ. Rev.* **53**(3), 433–443 (1963)
- Modigliani, F., Miller, M.H.: The cost of capital, corporation finance and the theory of investment. *Am. Econ. Rev.* **48**(3), 261–297 (1958)
- Mutalib, A.: Determinants of capital structure in cement industry: a case of Nigerian listed cement firms. *SSRN Electron. J.* (2011). <https://doi.org/10.2139/ssrn.1905096>
- Myers, S.C.: Determinants of corporate borrowing. *J. Financ. Econ.* **5**(2), 147–175 (1977)
- Myers, S.C., Majluf, N.S.: Corporate financing and investment decisions when firms have information that investors do not have. *J. Financ. Econ.* **13**(2), 187–221 (1984)
- Quynh, N.T., Luan, L.D., Vinh, L.H.: Factors affecting capital structure of enterprises listed on Ho Chi Minh stock exchange. *Bank. Sci. Training Rev.* (222) (2020)
- Nguyen, T.D.K., Ramachandran, N.: Capital structure in small and medium-sized enterprises: the case of Vietnam. *ASEAN Econ. Bull.* **23**(2), 192–211 (2006)
- Banchuenvijit, W.: the short - run relation between the Thai stock market and some other stock markets. *Int. J. Bus. Econ.* **1**(2), 9–16 (2009)



# Determinant of Capital Adequacy Ratio: Evidence from Commercial Banks in Vietnam

Nguyen Thi Nhu Quynh<sup>(✉)</sup> and Nguyen Duc Trung

Ho Chi Minh University of Banking, No. 36 Ton That Dam Street, Nguyen Thai Binh Ward,  
District 1, Ho Chi Minh City 700000, Vietnam  
{quynhntn, trungnd}@buh.edu.vn

**Abstract.** The Capital Adequacy Ratio (CAR) is an important measure indicating the level of safety in business operation activities. The paper is conducted to investigate the factors affecting banks' CAR in Vietnam during the period from 2008–2021. By using the sample data of 21 commercial banks with Bayesian mixed-effect regression, the results confirm that both bank-characteristic, macro-economic factors affect bank's capital adequacy ratio. The variables that have a strong positive impact on capital adequacy ratio include the ratio of equity to total assets, deposits ratio, liquidity ratio, CPI. The variables of bank size, profitability indicator, COVID-19 pandemic have strong negative impact. The ratio of loan to total assets, loan growth rate, CPI have a weak impact on CAR. From the results, the paper suggests some recommendations to increase this ratio in the future time.

**Keywords:** Bayesian mixed-effects · capital adequacy ratio · COVID-19

## 1 Introduction

The capital adequacy ratio (CAR) is an economic indicator that reflects the relationship between equity and risk-adjusted assets of commercial banks. This indicator is an important measure indicating the level of safety in business operation activities, which is built and developed by leading experts in the banking sector under the Basel Committee. In Vietnam, these days, according to circular No. 41/2016/TT-NHNN dated December 30, 2016, on prescribing the capital adequacy ratio for operations of banks and /or foreign bank branches of the Governor of the State Bank, commercial banks must maintain a capital adequacy ratio of at least 8%. Although Circular 41 is only “covered” in part by the Basel II Accord, ensuring these fairly stringent requirements necessitates significant efforts on the part of banks. According to SBV (2022), the average capital adequacy rate of banks in Vietnam is 11.59%, separately with the group of joint stock commercial banks, this rate is at the level 12.03%.<sup>1</sup> However, this ratio in Vietnam remains very low

<sup>1</sup> Access from [https://www.sbv.gov.vn/webcenter/portal/vi/menu/trangchu/tk/hdchtctctd/tkm\\_sctcb?\\_afzLoop=63880681920440224#%40%3F\\_afzLoop%3D63880681920440224%26centerWidth%3D80%2525%26leftWidth%3D20%2525%26rightWidth%3D0%2525%26showFooter%3Dfalse%26showHeader%3Dfalse%26\\_adf.ctrl-state%3Dwew2kbbqq\\_4en](https://www.sbv.gov.vn/webcenter/portal/vi/menu/trangchu/tk/hdchtctctd/tkm_sctcb?_afzLoop=63880681920440224#%40%3F_afzLoop%3D63880681920440224%26centerWidth%3D80%2525%26leftWidth%3D20%2525%26rightWidth%3D0%2525%26showFooter%3Dfalse%26showHeader%3Dfalse%26_adf.ctrl-state%3Dwew2kbbqq_4en).

when compared to the ASEAN + 5 countries, which range from 16% to 24%. (World Bank, 2022). As a result, the increase in capital is weighing on Vietnam's joint stock commercial banks, particularly small-scale banks. Therefore, understanding the factors affecting the CAR is essential for commercial banks to take measures to increase this ratio in the coming time.

According to the authors' knowledge regarding this topic, although several studies have been conducted in Vietnam as well as other countries, such as Vo, Nguyen & Do (2014), Pham & Nguyen (2017), Vu and Dang (2020), Aktas, Bakin & Celik (2015), El-Ansary & Hafez (2015), etc. However, these research results are not conclusive. Moreover, these studies mostly use some traditional estimation methods such as pooled OLS, FEM, REM, GMM. When applying these methods, the results are based only on data without incorporating prior information (Ngọc Thạch, 2019), thus there is a limitation that various studies mention in that the data is not enough to represent the population. In addition, these studies mainly focus on analyzing the impact of internal factors on the capital adequacy ratio without emphasizing the role of external factors.

This paper examines the determinants of the capital adequacy ratio of commercial banks in Vietnam during the period from 2008–2021. The paper makes several contributions to the existing literature. Firstly, it provides empirical evidence on the factors affecting the capital adequacy ratio with sample data of 25 commercial banks in Vietnam. Secondly, one of the limitations that previous studies usually mentioned is that data is not population, since most of them used frequentist inference. Different from these studies, in order to overcome that issue, the paper applied the Bayesian approach, this is a new point to supplement the research gap. Compare to frequentist inference, the Bayesian framework has several advantages, such as this method is based not only on research data but also on prior information, hence, with this combining, the results of Bayesian approach are more accurate as well as overcoming the limitation of data sample size. The third is, besides the bank-specific factors, this paper also implement several factors belonging to the macroeconomy, such as GDP growth, inflation and COVID-19 pandemic.

The remaining parts of this research are structured as follows. Section 2 presents the literature review, Sect. 3 describes the data, model, and methodology. Section 4 analyzes the empirical results and finally, we have some conclusions and suggest some policy implications.

## 2 Literature Review

Until now, there are pretty much previous studies that have explored the determinants of capital adequacy ratio. Aktas et al. (2015) analyze the impact of bank-dimensional and environmental factors on banks' capital adequacy ratio in the South Eastern European (SEE) region. Using the annual data from 71 commercial banks in 10 different countries in the SEE region in the time of 2007–2012, with a feasible GLS regression, the results show that both dimensional explanatory variables (such as bank size, ROA, leverage, liquidity, net interest margin, risk) and the environmental factors (such as economic growth rate, Eurozone stock market volatility index, deposit insurance coverage, and governance) have significant effects in determining CAR for the bank in the region.

When investigating Egyptian bank sectors, El-Ansary & Hafez (2015) use the sample data of 36 banks in the period from 2004–2013. The research results vary according to the period under study. In 2007–2013, liquidity, management quality and size are the most significant variable. Before the financial crisis, the variables affecting the capital adequacy ratio include size, asset quality and profitability. After 2009, asset quality, size, management quality, liquidity and credit risk are the most significant variables. Also, on this topic, El-Ansary & Hafez (2015) conduct research on the banking system in Albania. Using quarterly data from Q1/2007 to Q3/2014 of 16 private banks examine the factors affecting the capital. The results find that profitability indicators (such as ROA, ROE) do not have any influence on CAR, while non-performance loan, loan to deposit ratio and equity multiplier have a negative impact, whereas the bank size has a positive impact on capital adequacy ratio. In addition to these studies, several previous also consider the factors that influence on capital adequacy ratio, for example, Yu (2000), Bateni et al (2014), Ahmet & Hasan (2011), Kleff, V., & Weber, M. (2008).

Recently, Smaoui, Salah & Diallo (2020) have researched of determinants of capital in the Islamic banking system. By using a sample data of 122 Islamic banks during 2000–2014, the paper applies the system Generalized Method of Moments (GMM) estimator. The results indicate that deposit structure, bank size, and bank competition are significantly negatively related to Islamic banks' capital ratio, thus the authors confirm the "too-big-to-fail" theory. Besides, the generous deposits insurance system leads to lower Islamic banks' capital ratio.

Regarding the data sample of Vietnamese commercial banks, so far there have been a number of studies discussing the factors affecting the capital adequacy ratio, for example, Vo, et al. (2014); Pham & Nguyen (2017), Vu and Dang (2020),... In which, Vo et al. (2014) use the data from 28 commercial banks over a five-year period from 2007 to 2012. By FGLS regression, the research results find that liquid asset, and loan loss reserves have a positive impact on the capital adequacy ratio, whereas bank size, customer deposit ratio, and return on equity have a negative influence on the bank's CAR. With the sample data of 29 commercial banks in the period from 2011–2015, Pham and Nguyen (2017) use the fixed effect model (FEM) and random effect model (REM), the research results indicate that the ratio of net interest margin and liquidity ratio have a significant positive effect on CAR. But bank size and bank leverage (represented by ratio of equity to total liabilities) do not appear to have a significant effect on CAR. Variables loan loss reserves and loan to total assets are negatively related to CAR. Vu & Dang (2020) use data from 31 commercial banks during the period from 2011–2018. The results confirm that bank leverage, loan loss reserves, and return on equity have a negative impact, return on assets has a positive impact, while bank size, deposit, loan ratio, liquidity, net interest margin and non-performing loans do not significantly influence the CAR of Vietnamese commercial banks.

In sum, the topic of the determinant of capital adequacy ratio has received the attention of many scholars. However, these research results are inconclusive. In addition, most of the previous studies mainly study the internal factors without the macroeconomic condition. That is why the research is conducted to confirm the factors affecting on capital adequacy ratio of commercial banks in Vietnam, including the bank-specific and macroeconomic factors as well as the situation of the COVID-19 pandemic. From

the research results, the paper suggests some recommendations to increase the bank's CAR in the future time.

### 3 Method, Model and Data

#### 3.1 Methodology

To analyze the factors affecting on capital adequacy ratio, a multilevel model is applied. In literature, there are various sectors using multilevel approaches, from health, and social science to econometrics, for example, Simons-Morton, Simons Morton & Bunker (1988), who analyzed influencing personal and environmental conditions for community health; or Tseloni (2006), who investigated the impact of household and area on the incidence of total burglaries, property crimes, and thefts. In this paper, the authors use the multilevel (mixed-effects) perspective in a Bayesian approach for several reasons as follows.

Firstly, mixed-effects models are characterized as including random effects and fixed effects. The fixed effects model indicates that the individual-specific effect is correlated to independent variables, they are estimated directly as well as similar to standard regression parameters. The random effects are not estimated directly but are summarized according to their estimated variances and covariances. Random effects might take the form of either random intercepts or random parameters, and the grouping structure of the data may consist of multiple levels of nested groups. Hence, mixed-effect models are also known as hierarchical and multilevel models. Essentially, fixed effects are defined as the predictor variables which effects you are interested in after calculating for random variability (so, fixed). Random effects are as noise in the data. These are effects that arise from uncontrollable variability within the sample. Subject level variability is often a random effect.

Secondly, these days, the Bayesian probabilistic approach is sound more popular than traditional statistics. According to Nguyen & Nguyen (2018). Nguyen et al. (2019), the Bayesian framework has several strong advantages over traditional inference. Firstly, with the frequentist inference through some traditional estimations (such as pooled-OLS, fixed effect models, random effect models), these estimators were based on the data without incorporating prior information (Ngọc Thạch, 2019). Unlike this inference, the results of the Bayesian approach are based not only on data but also on prior observations. Thank this combination, the results of Bayesian inference are more accurate and reliable as a limit is not set to the data sample size. Besides, the traditional estimators have the possibility of omitting variable that is not significant despite potentially affecting the analysis, while the Bayesian approach considers the influence of all variables.

In this study, in order to investigate the determinants of capital adequacy ratio, this research employs a Bayesian mixed-effects regression, in which both two models without and with random effects (intercepts) are estimated. The difference between commercial banks in the initial capital adequacy ratio is reflected by random intercepts. The authors use GDP, CPI and DUMMY as three control variables in the research model. A Bayes factor test and a model test will be used to choose the more appropriate model. In terms of prior information, Lemoine (2019) strongly proposes informative priors, and Block et al. (2011; 2012) recommend standard Gaussian distributions for model parameters.

### 3.2 Description of Variable

#### 3.2.1 The Independent Variable: Bank Capital Adequacy Ratio (CAR)

In order to provide general principles and banking laws, the Basel Committee on Banking Supervision (BCBS) has proposed versions that commercial banks must comply with. In which, according to Basel 1, the CAR is calculated as follows:

$$CAR = \frac{\text{Capital (Tier 1 + Tier 2)}}{\text{Risk weighted assets (RWA)}}$$

According to Basel 1, banks are required to maintain this ratio at a minimum of 8%. Basel 1 divided the bank's equity into two categories: core capital and supplementary capital. Tier 1 capital is core capital, including permanent shareholders' equity and disclosed reserves. Tier 2 capital, is supplementary capital, including undisclosed reserves, asset revaluation reserves, general loan-loss reserves, and hybrid capital instruments. Tier 3 capital including short-term subordinated debt, this capital is not taken into account when calculating the capital adequacy ratio because of its lowest reliability.

The publication of Basel 1 along with detailed regulations had great significance for the risk management of commercial banks. However, the development of banking activities in the world had made the application of Basel 1 reveal several limitations. Hence, in 2004, the BCBS published Basel 2 guidelines aiming to refine and reform the version of Basel 1. Basel 2 is divided into three pillars related to minimum capital requirements, supervisory review and market discipline. In the first pillars, the CAR is set at a minimum of 8% and is calculated as follows:

$$CAR = \frac{\text{Capital}}{\text{RWA (Credit risk)} + \text{RWA (Operational risk)} + \text{RWA (Market risk)}}$$

These days, in Vietnam, the State Bank issued Circular No. 41/2016/TT-NHNN. Accordingly, the commercial banks must maintain a CAR of at least 8% and the CAR is calculated by the following formula. This formula is used to measure CAR in this paper:

$$CAR = \frac{C}{RWA + 12,5 (K_{OR} + K_{MR})} \times 100\% \quad (1)$$

In which:

C: Total equity capital

RWA: Credit risk adjusted Assets

$K_{OR}$ : Regulatory capital for operational risk

$K_{MR}$ : Regulatory capital for market risk

#### 3.2.2 The Factors Affecting Capital Adequacy Ratio

##### Internal Factors

##### Bank Size

According to Pham & Nguyen (2017), the logarithm of total assets is used to measure bank size. The literature on the banking sector shows that banks with larger scales have a

better reputation as well as are more experienced (Smaoui, Salah & Diallo, 2020). Hence, larger banks are easily able to diversify their asset portfolio. At the same time, many methods of mobilization are also implemented, which increases the capital adequacy ratio and reduces the risks. When analyzing the Albanian banking system in the period from 2007–2014, Shingjergji & Hyseni (2015) indicate that the positive relationship between bank size and CAR.

However, contrary to the above view, the theory of “too big to fail” argues that large banks typically hold a diverse portfolio of deposit claims, making their deposits less risky than those of small banks (Yu, 2000). This causes large banks to tend to take on excessive risk by allocating more capital to risky assets, in order to increase expected return, leading to increased risks for their assets portfolio. Several previous empirical pieces of evidence have shown the negative relationship between bank size and capital adequacy ratio, such as those of Dreca (2013), Bateni et al (2014), El-Ansary & Hafez (2015), and Akta et al. (2015).

### **Bank Leverage (LEV)**

In this paper, the authors measure bank leverage with the equity to total assets ratio. Hence, a high LEV denotes high equity or low leverage whereas a low LEV indicates low equity or high leverage. According to Ahmet & Hasan (2011), shareholders would discover that highly leveraged banks (lower equity to total assets ratio) are riskier than other banks. So they require a Nguyen Thi Nhu Quynh higher expected rate of return. Consequently, the high leveraged banks (lower equity to total assets ratio) may hold less equity capital and deal with difficulty in raising new equity because of the high cost of equity. So the authors suggest a positive correlation between LEV and capital adequacy ratio.

### **Loan Loss Provision**

Loan loss provision is calculated by the ratio of loan loss provision to total loan outstanding (Vo, et al. 2014). Vu and Dang (2020) indicate that the ratio of loan loss provision is a proxy for bank risk and this indicator could demonstrate the bank’s financial health. When a bank has a bad loan, it must set aside reserves, these provisions are taken from its earnings or its equity if earnings are not enough, which would reduce its capital. In addition, a higher loan loss provision ratio also indicates a higher bank risk, which would make it has more challenge to raise capital. So the paper believes that a negative link between the loan loss provision ratio and the capital adequacy ratio. In the literature term, this is consistent with several empirical evidence such as Aktas et al. (2015), Vu & Dang (2020), El-Ansary & Hafez (2015).

### **Deposit Ratio**

The deposit ratio is measured by the ratio of the customer deposit to total assets (Vo et al. 2014). According to Kleff & Weber (2008), comparing other sources of capital (new equity, bond financing), deposits can be the cheapest. In practical terms, customers tend to deposit at financial institutions with a good reputation and high brand, so an increase in deposits signals that banks and other financial institutions as financial intermediaries have implemented suitable capital mobilization strategies, and their brand is affirmed through the trust of a customer. The results of Masood & Ansari (2016) show a positive

relationship between deposit and capital adequacy ratio. Hence this paper is also expected to have a positive linking between deposit and CAR.

### **Loan Ratio**

The loan ratio is calculated as the ratio of the customer loan to total assets. On the balance sheet, total loans play the most important role in generating income for the bank. However, lending has two faces. Firstly, it provides the major earning for banks, and otherwise, lending is a source of credit risk. The credit risk and earnings from lending depend on the characteristics of the loan and the level of portfolio diversification of a bank. According to Vu and Dang (2020), the more loans extended, the higher the risk. Hence, a larger amount of capital will be needed to hedge the risk, so the research demonstrates the positive association between loan ratio and capital adequacy. This relationship is consistent with Mpuga (2002), and this is a reason why the paper expected a positive relationship between loan ratio and CAR.

### **Liquidity Ratio**

Most previous studies also agree that the relationship between liquidity ratio and CAR is positive. According to Bitar, Hassan & Hippler (2017), the banks with higher liquid assets are more able to raise debt, which could increase bank capital holdings. Angbazo (1997) states that as the proportion of funds invested in cash or cash equivalents increases, a bank's liquidity risk declines, leading to a lower liquidity premium in the net interest margins. Moreover, a higher level of bank liquidity has a favorable effect on the capital ratio by altering the required rate of return on bank shares (Mehranfar, 2013). When the bank ensures the input and output cash flow, it means ensuring liquidity, thereby helping the bank increase profits and capital sources. The CAR improves as a result of this. Hence, the authors suggest a positive associate between liquidity ratio and CAR in commercial banks in Vietnam.

### **Profitability**

According to the pecking order theory, firms in general and commercial banks, in particular, prefer internal over external financing (Rahman, 2019), internal capital can be mentioned as retained earnings. The reason is external financing emits various negative signals (Belkhir, Maghyreh & Awartani, 2016). Hence, when banks make a profit, they tend to use this profit to increase capital with the goal of making more earnings in the future. Several empirical pieces of evidence, such as El-Ansary & Hafez (2015), Keqa (2021) find a positive relationship between profitability and the capital adequacy ratio. As a result, the predicted sign of the profitability variable's coefficient is positive. There is a different proxy for profitability, similar to Vo et al. (2014), in this paper, the authors use an indicator of ROE to represent profitability.

### **Loan Growth Rate**

Banks are financial intermediaries, having the role of moving the capital from places of excess to places of shortage capital. Therefore, if loans increase, this could result in an increase in capital requirement (Ayuso, Pezer & Saurina, 2004). Thus, the relationship between loan growth rate and capital adequacy ratio is expected to be positive.



Besides internal factors, several external factors also influence capital adequacy ratios, such as gross domestic growth, inflation, or the COVID-19 pandemic situation. The next part of this paper will discuss these factors.

## External Factors

### Gross Domestic Product

When the economy has a good growth rate, investment activities, as well as the production and business of enterprises, are promoted. As a result, the bank's lending increases and encourages it to raise bank capital holdings. Moreover, according to Vithessonthi (2014), during economic booms, banks may increase their capital holding because of the rapid expansion of credit growth. So the predicted sign of this variable's coefficient can be positive.

### Inflation

According to Bitar et al. (2017), when the inflation of the economy is high, central banks will take some necessary measures to deal with this situation such as increasing interest rates, and increasing the required reserve, ... Thereby inducing firms and banks to borrow less, which favors the use of equity financing, so the paper expects a positive association between inflation and capital adequacy ratio.

### COVID-19

According to Özlem Dursun-de Neef & Alexander Schandlbauer (2022), during the COVID-19 pandemic, individuals and households were not able to spend money on relaxation activities because of mobility restrictions. As a result, they can accumulate savings in their deposit accounts. Hence, banks can increase deposits, and they also use additional funds to issue more real estate loans. This leads to an increase the bank capital holding, this is a reason why the paper suggests the positive link between COVID-19 pandemic and the capital adequacy ratio.

### 3.2.3 Specific Model

To examine the factors affecting on capital adequacy ratio in Vietnamese joint stock commercial banks, according to Smaoui et al. (2020), Vu & Dang (2020), Aktas. Bakin & Celik (2015), Vo et al. (2014), Ho & Hsu (2010) the study estimates a regression equation as follows:

$$\begin{aligned} CAR_{i,t} = & \alpha_0 + \alpha_1 BANKSIZE_{i,t} + \alpha_2 LEV_{i,t} + \alpha_3 LLR_{i,t} + \alpha_4 DEP_{i,t} + \alpha_5 LTA_{i,t} \\ & + \alpha_6 LIQ_{i,t} + \alpha_7 ROE_{i,t} + \alpha_8 LGR_{i,t} + \alpha_9 GDP_t \\ & + \alpha_{10} CPI_t + \alpha_{11} DUMMY_t + \mu_i + \epsilon_{i,t} \end{aligned} \quad (2)$$

where  $i$  and  $t$  refer to bank and year, respectively;  $\alpha_0$  is the constant,  $\mu_i$  and  $\epsilon_{i,t}$  are banks and time fixed effect.

Table 1 presents definition and the measurement and expected signs of the regression coefficients of the variable in research model.

**Table 1.** Definition and variables measurement in research model

Notation	Name of variables	Measure/Data source	Sign of expectation	Researches
<b>Dependent variable</b>				
CAR	Capital adequacy ratio	$CAR = \frac{C}{RWA + 12,5(K_{OR} + K_{MR})}$		Vu and Dang (2020), Nguyen & Pham (2017), Shingjergji & Hyseni (2015), Vo et al. (2014)
BANKSIZE	Bank size	Logarithm (Total asset)	±	Smaoui et al. (2020), Dreca (2013), Bateni et al (2014), El-Ansary & Hafez (2015), Akta et al. (2015), Shingjergji & Hyseni (2015)
LEV	Bank leverage	$\frac{\text{Equity}}{\text{Total assets}}$	+	Smaoui et al. (2020), Ahmet & Hasan (2011), Vo et al. (2014), Vu and Dang (2020)
LLR	Loan loss provision	$\frac{\text{Loan loss provision}}{\text{Total loan outstanding}}$	–	Vo et al. (2014), Vu and Dang (2020), Aktas et al.(2015), El-Ansary &Hafez (2015)
DEP	Deposit ratio	$\frac{\text{Customer deposit}}{\text{Total assets}}$	–	Vo et al. (2014), Kleff & Weber (2008), Masood & Ansari (2016)
LTA	Loan ratio	$\frac{\text{Customer loans}}{\text{Total assets}}$	+	Vu and Dang (2020), Nguyen & Pham (2017)

*(continued)*

**Table 1.** (continued)

Notation	Name of variables	Measure/Data source	Sign of expectation	Researches
Dependent variable				
LIQ	Liquidity ratio	$\frac{\text{Cash and Cash Equivalents}}{\text{Total assets}}$	+	Aspal & Nazneen (2014); El-Ansary & Hafez (2015), Akta et al. (2015),
ROE	Profitability	$\text{ROE} = \frac{\text{Net income}}{\text{Equity}}$	+	El-Ansary & Hafez (2015), Keqa (2021)
LGR	Loan growth rate	$\frac{\text{Cash and Cash Equivalents}}{\text{Total assets}}$	+	Vo et al. (2014), Smaoui et al. (2021)
GDP	Growth domestic product	$\frac{\text{GDP}_t - \text{GDP}_{t-1}}{\text{GDP}_{t-1}}$	+	Akta et al. (2015), Smaoui et al. (2021)
CPI	Inflation	$\frac{\text{CPI}_t - \text{CPI}_{t-1}}{\text{CPI}_{t-1}}$	+	Smaoui et al. (2021)
DUMMY	COVID-19	Dummy variable with a value of 1 in year of the COVID-19 pandemic and 0 for the remaining years	+	Suggested by the authors

Source: Various authors

### 3.3 Data Description

This paper uses data from 25 commercial banks in Vietnam during the period from 2008–2021. According to the State Bank of Vietnam (SBV) (2022), there are 31 Vietnamese joint stock commercial banks, but some of them are data omissions. This is the reason why our database includes 25 commercial banks. To examine the factors affecting on capital adequacy ratio, the authors use both bank-level and country-level data. In which, the bank-level data is taken from the audited financial statement or annual report whereas the country-level data is derived from the database of the World Bank. Regarding the COVID-19 variable (DUMMY), it has a value of 1 in the years of the COVID-19 pandemic (2020 and 2021) and 0 for the remaining years.

**Table 2.** Descriptive statistics of variables

Variable	Obs	Mean	Std. Dev	Min	Max
CAR	329	0.136	0.050	0.066	0.459
BANKSIZE	329	8.015	0.539	6.470	9.250
LEV	329	0.096	0.045	0.041	0.356
LLR	329	0.013	0.006	0.000	0.040
DEP	329	0.638	0.122	0.292	0.894
LTA	329	0.560	0.130	0.194	0.852
LIQ	329	0.013	0.015	0.002	0.124
ROE	329	0.114	0.076	0.000	0.315
LGR	329	0.246	0.252	-0.313	1.650
GDP	329	0.059	0.014	0.026	0.072
CPI	329	0.065	0.059	0.006	0.231
DUMMY	329	0.140	0.347	0.000	1.000

*Source: The authors' calculations*

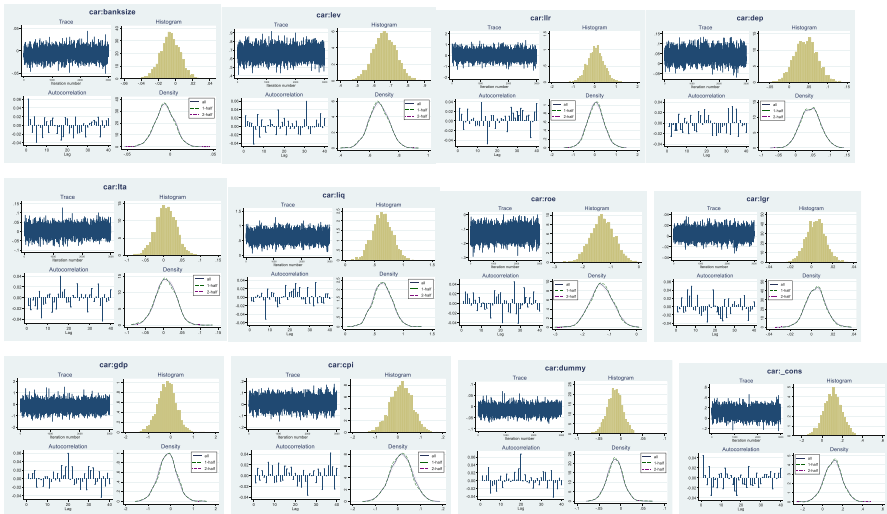
The results of descriptive statistics of variables in research models are summarized in Table 2 with unbalanced panel data. The observations for each of the variables are 329. The average CAR is 13.6%, which is higher than the minimum ratio prescribed by the Basel committee and SBV. The lowest CAR is 6.62%, which is to the Bank for Investment and Development of Vietnam (BIDV) in 2009. The highest value is 45.89%, which belongs to Eximbank in 2009. Regarding independent variables, the mean value of bank size is 8.015 with the highest and the lowest bank sizes at 6.47 and 9.25, respectively. This indicates that the commercial banks in Vietnam are diverse in scale. The average LEV is 9.6%, which shows that the bank's assets mainly come from liabilities. The average LLR is 1.3%, which signals the proportion is quite low. The mean values of DEP and LTA are 63.8% and 56.0%. Respectively. These indicate that customer deposits are important financing for banks. At the same time, customer loans are mainly banks' assets. The LIQ variables have an average value of 1.3%, showing that the ratio of cash reserves for commercial banks is quite low, which creates motivation to increase earnings. The mean value of the loan growth rate is 24.6%, showing that the loan growth rate for customers of Vietnamese commercial banks is at an average level. In the period from 2008–2021, the mean value of CPI is higher than GDP, which indicates that for several years the economy is not efficient due to the impact of the financial crisis as well as the COVID-19 pandemic.

## 4 Bayesian Simulation Results

### 4.1 MCMC Convergence Diagnostics

In order to test the validity of Bayesian inference, we need to check the MCMC convergence, efficiency, and acceptance rate. In which, the mixing properties of MCMC are indicated by efficiency rate. Efficiency indicates the mixing properties of MCMC sequences. High-efficiency rate shows that MCMC sequences mix well, whereas low efficiency implies bad mixing in the simulated MCMC sample. In model research, the acceptance rate obtains 0.81 (the required rate is 0.1), The min, average and max of efficiency rates are 0.92; 0.98 and 1, respectively (the required rate is 0.01). Thus, regarding the acceptance rate and efficiency rate, these rates are satisfied for Bayesian inference.

Concerning the test for chain convergence. The results of chain convergence are presented in Fig. 1. From Fig. 1, the chain convergence, including traces, autocorrelation, cusum and histogram plots show no convergence issue. In particular, the trace plots traverse quickly through the posterior domain, exhibiting no trends; the autocorrelations have no lags; the cusum plots are jagged, intercepting the X axis; the histogram plots resemble the shape of the posterior distributions of the model parameters. To summarize, we can conclude that the parameters of our model have converged to reasonable values.



**Fig. 1.** Convergence diagnostics for the model parameters. *Source: The authors' calculations*

Table 3 denotes that all the parameters of the model have an efficiency of more than 0.91, while the warning level is 0.1. Furthermore, all the correlation times are relatively small.

Thus, from the results of chain convergence, the acceptance and efficiency rates as well as Effective sample size, we can conclude that MCMC sequences have converged to the desired distribution and we can proceed to inference.

**Table 3.** Effective sample size

	ESS	Corr. time	Efficiency
CAR			
BANKSIZE	2785.45	1.08	0.9285
LEV	2796.79	1.07	0.9323
LLR	2869.24	1.05	0.9564
DEP	3000.00	1.00	1.0000
LTA	3000.00	1.00	1.0000
LIQ	3000.00	1.00	1.0000
ROE	3000.00	1.00	1.0000
LGR	3000.00	1.00	1.0000
GDP	2938.64	1.02	0.9795
CPI	2999.22	1.00	0.9997
DUMMY	2930.67	1.02	0.9769
_cons	2754.99	1.09	0.9183
sigma2	2927.19	1.02	0.9757

Source: The authors' calculations

## 4.2 Interpretation of Empirical Results

All the model parameters are summarized in Table 4. From Table 4, Monte Carlo chain standard error (MCSE) estimates are close to zero, which indicates that the MCMC algorithm is reasonable. In general, the estimate will have a higher accuracy when the MCMC is lower. Unlike frequentist inference, in Bayesian inference, 95% credible intervals indicate which range the true value of a certain parameter belongs to, e.g., the mean value of variable BANKSIZE lies in an interval between  $-0.0294$  and  $0.0163$  with a 95% probability, and so on.

In view of probability, variables that have a positive effect on capital adequacy ratio include bank leverage (LEV), loan loss provision (LLP), customer deposit ratio (DEP), loan to total assets ratio, liquidity ratio (LIQ), loan growth rate (LGR) and inflation (CPI). In which, the variables of bank leverage (LEV), customer deposit ratio (DEP), liquidity ratio (LIQ) strongly positively contributes to the capital adequacy ratio. The variables of bank size (BANKSIZE), profitability (ROE), growth domestic product (GDP) and COVID-19 pandemic (DUMMY) have a strong negative impact on CAR.

Based on the empirical results, the paper has the following discussion.

Firstly, the regression coefficient of variable BANKSIZE is negative with 71.07% probability. That means the larger the bank size, the lower the capital adequacy ratio, this result is similar to Usman, Lestari & Puspa (2017) when they use the data sample for the banking sector in Indonesia, or Bateni, Vakilifard, & Asghari (2014) when they applied sample data for Iranian banks. According to Usman et al. (2017), a larger size bank usually has a smaller risk, as a reason, the capital adequacy ratio is not as high

**Table 4.** Posterior model summary

	Mean	Std. Dev.	MCSE	Median	Probability of coefficient mean > 0	Equal-tailed [95% Cred. Interval]
CAR						
BANKSIZE	-0.0062	0.0116	0.0002	-0.0063	0.2893	[-0.0294, 0.0163]
LEV	0.6626	0.0687	0.0013	0.6625	1.0000	[0.5274, 0.7970]
LLR	0.0435	0.3839	0.0072	0.0399	0.5447	[-0.7084, 0.8161]
DEP	0.0427	0.0293	0.0005	0.0430	0.9323	[-0.0136, 0.1012]
LTA	0.0060	0.0275	0.0005	0.0062	0.5873	[-0.0489, 0.0579]
LIQ	0.6392	0.1662	0.0030	0.6386	1.0000	[0.3235, 0.9639]
ROE	-0.1299	0.0418	0.0008	-0.1305	0.0007	[-0.2101, -0.0489]
LGR	0.0040	0.0091	0.0002	0.0041	0.6633	[-0.0139, 0.0223]
GDP	-0.1516	0.4030	0.0074	-0.1450	0.3543	[-0.9564, 0.6314]
CPI	0.0145	0.0486	0.0009	0.0154	0.6230	[-0.0826, 0.1083]
DUMMY	-0.0136	0.0181	0.0003	-0.0134	0.2147	[-0.0497, 0.0215]
_cons	0.1066	0.0904	0.0017	0.1090	0.8827	[-0.0730, 0.2817]
sigma2	0.0012	0.0004	0.0000	0.0011		[0.0007, 0.0022]

Source: The authors' calculations

as a bank with a smaller scale. In practical terms, large banks usually have a high level of security, because they have large enough capital to bear any risky assets. Hence, the capital adequacy ratio has a negative impact on bank size.

Secondly, bank leverage (represented by the ratio of equity to total assets) has a strong positive on the capital adequacy ratio. This result is consistent with initial expectations as well as several previous studies, such as Usman et al. (2017), Ho & Hsu (2010). And this is also completely consistent with the fact that banks with high equity ratios will hold more equity capital, so they will easily raise capital. As a result, the LEV has a strong positive association with the capital adequacy ratio.

Thirdly, the regression coefficient of variable LLR is positive with 54.47% probability, showing that the loan loss provision has an ambiguous impact on the capital adequacy ratio. In fact, loan loss provisions are cash reserves set aside by a bank in anticipation of potential losses from lending (Vu & Dang, 2020). So, the ratio of loan loss provision to total loan is a proxy for bank risk. Therefore, on the one hand, the larger the provision, the more negative impact on the bank's earnings. However, on the other hand, the more provisioning, the more banks are lending. This creates an incentive to increase the bank's earnings, thereby leading to an increase in equity and capital adequacy ratio. So, the ambiguous link between loan loss provision and capital adequacy ratio is also acceptable.

Four is both deposits and loans to total assets have a positive effect on the capital adequacy ratio. In which, the variable of deposit ratio has a very strong impact while the variable of loan to total assets has a relatively weak impact. This is reflected in the reality of the business activities of commercial banks in Vietnam. In reality, customer deposits are the cheapest financing, which is the premise for the bank to perform other business operations and generate bank profit. Whereas lending activities have two opposite sides, on the one hand, lending activities will promote generating the main bank's income; on the other hand, in the case the banks have poor loan management efficiency, it will negatively affect loan quality, thereby increasing non-performing loans. As a result, the bank's profit would reduce. Therefore, customer deposit has a strong positive while the loan to total assets has a weak positive effect on the capital adequacy ratio.

Fifth, the results show that bank liquidity affects a very strong positive on the capital adequacy ratio (the regression coefficient is positive with 100 probability). This result is similar to initial expectations and most previous studies, such as Angbazo (1997), Aspal & Nazneen (2014), El-Ansary & Hafez (2015), Akta et al. (2015). Literally, when the ratio of cash or cash equivalents increases, the bank's liquidity is higher. As a result, the capital adequacy ratio also increases. So the link between liquidity ratio and CAR is a very strong positive.

Sixth, the regression coefficient of ROE (a proxy for profitability indicator) is negative with a probability of approximately 100%. Although this result contradicts the initial expectation as well as some research by Gropp & Heider (2007). However, this result is similar to Vo et al. (2014) when they analyze commercial banks in Vietnam or Jim Wong, Ka-fai Choi & Tom Fong (2005) with their database of banks in Hong Kong. This finding indicates that commercial banks usually try to achieve the goals of shareholder wealth maximization by deciding to invest as much as possible in profitable assets withholding capital from internal financing such as retained earnings. Then, the banks tend to invest in riskier portfolios and loans, which leads to increase bank risks and thus CAR decrease.

Seventh, the positive regression coefficient between the loan growth rate and the capital adequacy ratio with a probability of 66.33% signals the ambiguous relationship between the loan growth rate and CAR. This finding is consistent with the low-level study's sign expectation. The reason is that in Vietnam, commercial banks are not allowed to grow credit freely. Instead, they have to follow the control of the State Bank to match the macroeconomic situation.



Finally, the economic condition also affects the capital adequacy ratio. In which, GDP is negative whereas the inflation rate (CPI) is a positive effect. These harmonize perfectly with practicals in Vietnam. When the economy has a good growth rate, enterprises, and individuals tend to borrow to invest. At the same time, commercial banks also tend to use idle capital to make profitable investments, which leads to increase bank risk and reduced capital adequacy ratio. On contrary, when the economy has high inflation, State Bank has various measures to deal with the situation (Bitar et al., 2017). As a result, firms and individuals borrow less, which increases the bank holding capital, thereby increasing the capital adequacy ratio. Finally, the COVID-19 pandemic (represented by the DUMMY variable) has a negative influence on the capital adequacy ratio. Although this result is not consistent with the expectation of research, however, it reflects the actual situation in Vietnam. During the COVID-19 pandemic (the year 2020, 2021), due to the impact of social distancing, enterprises, households, and individuals reduced borrowing. But at the same time, a relatively large amount of capital was withdrawn from banks to invest in the stock market channel. This leads to reducing the holding capital of the banks, thus, their CAR is reduced.

In sum, among of variables, bank size, profitability indicators, and the COVID-19 pandemic have a strong negative impact, whereas bank leverage (represented by the ratio of equity to total assets), deposits ratio, liquidity ratio, and CPI are motivating factors for a bank to increase the CAR. The variables, ratio of loan to total assets, loan growth rate, and CPI have ambiguous influences on the dependent variable. From these findings, the next section of the paper suggests some recommendations to increase the capital adequacy ratio.

## 5 Conclusion

This research investigates the determinants of the capital adequacy ratio of commercial banks in Vietnam in the time of 2008–2021. By Bayesian mixed-effects regression with the sample data of 25 commercial banks, the results show that both external factors and internal factors affect CAR. In which, the variables that have a strong positive impact on the capital adequacy ratio include the ratio of equity to total assets, deposits ratio, liquidity ratio and CPI. Whereas, the variables of bank size, profitability indicator and COVID-19 pandemic have strong negative. The ratio of loan to total assets, loan growth rate and CPI have a weak impact on CAR. These findings are consistent with various previous studies such as Vo et al. (2014), Akta et al. (2015), Nguyen & Pham (2017). Thus, the paper has achieved its state objective, by using a Bayesian mixed-effect estimator and has overcome the limitation of previous studies as the sample data is not representative of the population.

From the above results, the paper suggests several important recommendations as follows: (1) the results indicate that bank expansion reduces the bank's capital adequacy ratio. Hence, the State Bank should control and supervise the process of expanding the bank's scale. In addition, State Bank should be flexible in requesting an increased CAR to avoid increasing the bank's risks; (2) increasing the ratio of equity to total assets increases the capital adequacy ratio, so the banks should consider distribution policy by increasing retained earnings to increase the bank's equity. Besides, in order

to increase equity in the future, commercial banks should increase the investment of strategic shareholders; (3) the results find that the customer deposits would promote the increase of capital adequacy ratio, so the commercial banks should have various preferential policies to encourage the customer to deposit, as well as increase advertising and marketing strategies; (4) in order to increase the capital adequacy ratio, one of the important implications is that commercial banks should increase assets with high liquidity by actively developing a framework policy on liquidity risk management; (5) regarding profitability indicator, the research results show that the return on equity has a strong negative effect on capital adequacy ratio. Hence, banks need to ensure that the implementation of increasing profitability must always be closely combined with the regulations on risk control in a reasonable and specific way; (6) in the context of a rapidly growing economy, commercial banks need to be alert in the process of building lending and investment strategies to avoid risks for banks; (7) And finally, the results indicate that during the CPVID-19 pandemic, the bank's capital adequacy ratio decreases. Although in this period, enterprises, households, and individuals all borrowed less, the bank's capital flows tended to flow out into other investment channels. For this reason, the State Bank and commercial banks should consider solutions such as marketing or increasing deposit interest rate to ensure a capital adequacy ratio.

## References

- Ahmet, B.Y.K., Hasan, A.: Determinants of capital adequacy ratio in Turkish banks: A panel data analysis. *Afr. J. Bus. Manage.* **5**(27), 11199–11209 (2011)
- Aktas, R., Bakin, B., Celik, G.: The determinants of banks' capital adequacy ratio: Some evidence from southeastern European countries. *J. Econ. Behav. Stud.* **7**(1(J)), 79–88 (2015)
- Angbazo, L.: Commercial bank net interest margins, default risk, interest-rate risk, and off-balance sheet banking. *J. Bank. Financ.* **21**(1), 55–87 (1997)
- Ayuso, J., Pérez, D., Saurina, J.: Are capital buffers pro-cyclical?: Evidence from Spanish panel data. *J. Financ. Intermediation* **13**(2), 249–264 (2004)
- Bateni, L., Vakilifard, H., Asghari, F.: The influential factors on capital adequacy ratio in Iranian banks. *Int. J. Econ. Financ.* **6**(11), 108–116 (2014). <https://doi.org/10.5539/ijef.v6n11p108>
- Belkhir, M., Maghyereh, A., Awartani, B.: Institutions and corporate capital structure in the MENA region. *Emerg. Mark. Rev.* **26**, 99–129 (2016)
- Bitar, M., Hassan, M.K., Hippler, W.J.: The determinants of Islamic bank capital decisions. *Emerg. Mark. Rev.* **35**, 48–68 (2018)
- Block, J.H., Jaskiewicz, P., Miller, D.: Ownership versus management effects on performance in family and founder companies: A Bayesian reconciliation. *J. Fam. Bus. Strat.* **2**(4), 232–245 (2011)
- Block, J.H., Hoogerheide, L., Thurik, R.: Are education and entrepreneurial income endogenous? A Bayesian analysis. *Entrepreneurship Res. J.* **2**(3) (2012)
- Dreca, N.: Determinants of capital adequacy ratio in selected Bosnian banks. *Dumlupinar Univ. J. Soc. Sci.* **12**(1), 149–162 (2014)
- El-Ansary, O., Hafez, H.: Determinants of capital adequacy ratio: An empirical study on Egyptian banks. *Corp. Ownersh. Control.* **13**(1), 806–816 (2015)
- Wong, J., Choi, K.-F., Fong, T.P.-W.: Determinants of the capital level of banks in Hong Kong. In: Genberg, H., Hui, C.-H. (eds.) *The Banking Sector in Hong Kong*, pp. 159–190. Palgrave Macmillan UK, London (2008)

- Gropp, R., Heider, F.: What can corporate finance say about banks' capital structures? Working paper (2007). <http://wiwi.unifranfurt.de/schwerpunkte/finance/master/brown/177.pdf>
- Dursun-de Neef, H.Ö., Schandlbauer, A.: COVID-19, bank deposits, and lending. *J. Empir. Financ.* **68**, 20–33 (2022)
- Ho, S.J., Hsu, S.C.: Leverage, performance and capital adequacy ratio in Taiwan's banking industry. *Jpn. World Econ.* **22**(4), 264–272 (2010)
- Keqa, F.: The determinants of banks' capital adequacy ratio: evidence from Western Balkan countries. *Bus. Econ. J.* **12**(3), 1–6 (2021)
- Kleff, V., Weber, M.: How do banks determine capital? Evidence from Germany. *German Econ. Rev.* **9**(3), 354–372 (2008)
- Lemoine, N.P.: Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos* **128**, 912–928 (2019)
- Masood, U., Ansari, S.: Determinants of capital adequacy ratio. A perspective from Pakistani banking sector. *Int. J. Econ. Commerce Manag.* **4**(7), 247–273 (2016)
- Mehranfar, M.: Investigating the impact of bank efficiency and macroeconomic variables on risk management of banks. *Int. J. Appl.* **1**(1), 37–42 (2013)
- Mpuga, P.: The 1998–99 banking crisis in Uganda: What was the role of the new capital requirements? *J. Financ. Regul. Compliance* **10**(3), 224–242 (2002)
- Nguyen, H.T., Thach, N.N.: A panorama of applied mathematical problems in economics. *Thai J. Math.* 1–20 (2018). Special Issue: Annual Meeting in Mathematics
- Nguyen, H.T., Trung, N.D., Thach, N.N.: Beyond traditional probabilistic methods in econometrics. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *ECONVN 2019*. SCI, vol. 809, pp. 3–21. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_1](https://doi.org/10.1007/978-3-030-04200-4_1)
- Ngọc, T.N.: Một cách tiếp cận Bayes trong dự báo tổng sản phẩm quốc nội của Mỹ. *Tạp chí Kinh tế và Ngân hàng châu Á* **163**, 5–18 (2019)
- Pham, T.X.T., Nguyen, N.A.: The determinants of capital adequacy ratio: The case of the Vietnamese banking system in the period 2011–2015. *VNU J. Econ. Bus.* **33**(2) (2017)
- Rahman, M.T.: Testing trade-off and pecking order theories of capital structure: evidence and arguments. *Int. J. Econ. Financ. Issues* **9**(5), 63 (2019)
- Simons-Morton, D.G., Simons-Morton, B.G., Parcel, G.S., Bunker, J.F.: Influencing personal and environmental conditions for community health: a multilevel intervention model. *Fam. Community Health* **11**(2), 25–35 (1988)
- Shingjergji, A., Hyseni, M.: The determinants of the capital adequacy ratio in the Albanian banking system during 2007–2014. *Int. J. Econ. Commer. Manag.* **3**(1), 1–10 (2015)
- Smaoui, H., Salah, I.B., Diallo, B.: The determinants of capital ratios in Islamic banking. *Q. Rev. Econ. Financ.* **77**, 186–194 (2020)
- Tseloni, A.: Multilevel modelling of the number of property crimes: Household and area effects. *J. R. Stat. Soc. A. Stat. Soc.* **169**(2), 205–233 (2006)
- Vithessonthi, C.: The effect of financial market development on bank risk: Evidence from Southeast Asian countries. *Int. Rev. Financ. Anal.* **35**, 249–260 (2014)
- Vo, H.D., Nguyen, M.V., Do, T.T.: Yếu tố quyết định tỷ lệ an toàn vốn: Bằng chứng thực nghiệm từ hệ thống ngân hàng thương mại Việt Nam. *TẠP CHÍ KHOA HỌC ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH-KINH TẾ VÀ QUẢN TRỊ KINH DOANH* **9**(2), 87–100 (2014)
- Vu, H., Dang, N.: Determinants influencing capital adequacy ratio of Vietnamese commercial banks. *Accounting* **6**(5), 871–878 (2020)
- World Bank (2022). No time to waste: The challenges and opportunities of cleaner trade for Vietnam. Access from [https://www.worldbank.org/vi/country/vietnam/publication/taking-stock-vietnam-economic-update-january-2022?cid=eap\\_fb\\_vietnam\\_vn\\_ext&fbclid=IwAR3YWiLlzm0Bxyv55iYJn17AtD7PbF4OI4P7ZqevGxuLS1q9URxWsv5GxT0](https://www.worldbank.org/vi/country/vietnam/publication/taking-stock-vietnam-economic-update-january-2022?cid=eap_fb_vietnam_vn_ext&fbclid=IwAR3YWiLlzm0Bxyv55iYJn17AtD7PbF4OI4P7ZqevGxuLS1q9URxWsv5GxT0)
- Yu, H.C.: Banks' capital structure and the liquid asset- policy implication of Taiwan. *Pac. Econ. Rev.* **5**(1), 109–114 (2000)



# Impact of Managers' Gender Difference on Firms' Liability in Vietnam

Van Tung Nguyen<sup>(✉)</sup>, Nhan Truong Thanh Dang, Van Dung Ha,  
and Thi Anh Tuyet Le

Banking University Hochiminh City, Ho Chi Minh City, Vietnam  
{tungnv, nhandtt, dunghv, tuyetlta}@buh.edu.vn

**Abstract.** The article investigates the impact of the gender difference of entrepreneurs on the debt ratio of small and medium enterprises in Vietnam. The data source is based on survey results from 2007 to 2015 with more than 2600 enterprises. Research results through Bayesian estimation method show that male entrepreneurs use a higher corporate debt ratio than women. The study also found that the size of the enterprise, the education and training level of the entrepreneurs, and the export factor have a positive impact while the age of the enterprise has a negative impact on the debt ratio of the enterprise.

**Keywords:** Manager · Gender difference · SMEs · Liability · Vietnam

## 1 Introduction

There has been an increasing academic attention to the topic of the effect of characteristics of shareholders and managers on various variables related to enterprises. Examples of some most analyzed characteristics include the shareholder structure (Weisbach, 1988), the arrangement of the board of directors, and the combination of the positions of CEO and chairman of the board of directors in the same person (Boyd, 1994).

The managers' gender difference has been one of the characteristics attracting interest of researchers, for example: the research topic of the impact of presence of women on corporate boards of directors on firms' financial performance (Terjesen et al., 2009). The presence of women on the board has become a remarkable focus, not only for academic reasons but also for social concerns. In recent years there has been more pressure from society to allocate women to positions within the boards of directors. Based on the research of Soltane (2009), more women occupying management positions would be better for firms' performance.

There is still limited research on financial decisions and their correlation with the gender of managers, especially within the field of small enterprises (Hernandez et al., 2015). The contribution of women as managers or directors in small and medium enterprises has been increasingly acknowledged, given that such firms' governing bodies tend to be smaller, less structured and less complicated, indicating that each female manager has a greater capacity to influence decision-making process (Judge & Zeithaml, 1992).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 498–512, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_35](https://doi.org/10.1007/978-3-031-35763-3_35)

Gender may have significant impact on the level of corporate liability, as well as its cost and maturity (Hernandez et al., 2015). There are two main reasons why this should be the case. Firstly, women often tend to accept lower levels of risk (Smith et al., 2006). On the other hand, there may be gender discrimination on the part of the credit supply (Brush, 1992).

Within the Vietnam context, Vietnamese women have been allocated to leading positions on the business map compared to other countries in the region, according to Boston Consulting Group (BCG) (Vnexpress, 2018). Women held 25 percent of CEO or board-level positions in Vietnam, compared to 14 percent in Malaysia, 10 percent in Singapore, and only 6 percent in Indonesia, as cited by Bloomberg (Vnexpress, 2018). Based on a report by Deloitte in June 2017, Vietnam performed much better than its Asian peers, with women making up 17.6 percent of its corporate boards, higher than in Malaysia and Singapore. However, Vietnam still suffers the most from gender inequality, with a low female-to-male ratio in top management (Vnexpress, 2018). Little, if any, research has been conducted into the impact of manager gender differences on financial decisions in general and liability levels in firms in particular within the context of a country.

Considering the highlighted literature gap, this study aims to evaluate the impact of gender differences in management composition on firm liability in Vietnam, with a focus on SMEs across different industries in the country.

Theoretically, this research can contribute to the development of the literature on the topic of gender differences' impact on firms' financial performance in general and firms' liability in particular, with a focus on SMEs in a developing country like Vietnam. This empirical research can be viewed as a response to the call raised by previous researchers for more studies on the causal relationship between gender diversity and firms' liability across different regions around the globe (Novailitis et al., 2006; Romani et al., 2012; OECD 2013; Hernandez et al., 2015). This study can also be a considerable reference for academies having an interest in the characteristics of managers in emerging contexts. Practically, this study can help to provide managerial implications for the purpose of enhancing better liability management, taking advantage of gender diversity in the management board, and eventually contributing to the effective financial performance of firms.

## 2 Literature Review

### *Theories explaining enterprises' financial capital structures.*

There have been non-excluding theories which attempt to explain existing firms' capital structures: trade-off theory and pecking order theory (Hernandez et al., 2015). The trade-off theory helps to explain that each firm has an optimal liability or debt level based on the balance between advantages and disadvantages of market imperfections (mainly taxes, costs of financial distress, and agency costs and information asymmetry).

Pecking order theory was first introduced by Myers and Majluf (1984), who considered that the empirical evidence was inconsistent with a financial policy determined by a tradeoff of the advantages and disadvantages of market imperfections. Instead, firms' financial policies seem to be better reasoned by the behaviour described by Donaldson

(1961). He created a hierarchy describing company preferences for internal financial sources over external financial sources. Nevertheless, trade-off and pecking-order were not the only relevant theories associated with capital structure. Another significant theory is the financial growth cycle theory (Berger & Udell, 1998). According to this theory, there are several factors having impact on the level of information asymmetry in companies, and the age and the size are amongst the most important ones.

***Previous research related to gender differences' impact on financial behaviors or financial decisions.***

There have been previous studies focusing in highlighting several factors that differentially influence men and women's financial decisions. Some of the most extensively identified determination factors in the literature regarding gender differences in financial behavior are risk attitudes (Croson & Gneezy, 2009; Eckel & Grossman, 2008), financial knowledge (Fonseca et al., 2012; Lusardi & Mitchell, 2011), overconfidence (Goldsmith & Goldsmith, 1997) and cognitive style (Sladek et al., 2010).

Several previous studies have focused on different risk preferences by men and women have been across a wide range of domains. According to a substantial body of research in many areas, females tend to be more averse to risk than males (Byrnes et al., 1999). Behavioral economics theory suggests that women are less prone to risk taking than men (Croson and Gneezy 2009; Charness and Gneezy 2012). A similar pattern of higher willingness for risk taking in men, as compared to women, was also found by many studies in the field of investment and gambling (Berggren & Gonzalez, 2010; Croson & Gneezy, 2009; Eckel & Grossman, 2008; Kunnanatt & Emiline, 2012). Women with lower levels of risk tolerance tend to have less risky asset portfolios, exhibit greater relative risk aversion in their distribution of wealth and demonstrate lower willingness to accept financial risk as compared with men (Arano, Parker & Terry, 2010; Halko, Kaustia & Alanko, 2012; Siva, 2012).

Despite the wide acknowledgement of gender related discrepancies in risk perception, there is still lack of harmony regarding the causes and their underlying mechanisms. Several studies have discussed contradictory evidence within particular contexts. For instances, Schubert, Brown, Gysler & Brachinger (1999) found gender dissimilarities in risk preferences when the tasks involved gambling decisions. Schubert et al. (1999) concluded that under specific circumstances men might be more risk-averse than women. Eckel & Grossman (2008) reviewed several studies examining discrepancies in risk-related behavior of men and women. They made a distinction between laboratory studies presenting abstract gambles, laboratory studies presenting contextual environments of investments or insurance decisions, and field studies evaluating men and women's actual gambling behavior and decisions of real-life investment. While the majority of the reviewed laboratory experiments found that women are less risk-tolerant than men, some studies involving investment and insurance frames presented counterevidence. Furthermore, some cross-cultural studies found gender differences among whites, but not among other ethnic groups (Feng & Seasholes, 2008). However, women seemed more risk averse than men in most countries (OECD, 2013).

Recently, financial literacy has been broadly recognized as a basic skill which is critical for individuals' financial decision making in today's modern complicated financial environment. An extensive number of empirical research provides evidence that

men have higher levels of financial literacy (Chen & Volpe, 2002; Fonseca et al., 2012; Lusardi & Mitchell, 2011; Zissimopoulos et al., 2008). For instances, Zissimopoulos et al., (2008) found that less than 20% of middle-aged college-educated women could solve a basic compound interest question compared to about 35% of college-educated males of the same age. Chen & Volpe (2002) found that women have lower financial knowledge than men. These patterns were gained across different countries (OECD, 2013), and a wide range of age groups (Agnew & Harrison, 2015; Chen & Volpe, 2002).

There has been a literature gap regarding knowledge about mechanisms leading to gender related disparities in financial literacy (Fonseca et al., 2012). Few studies on the mechanisms which underlie the observed gender gap provide some insights on possible influential factors such as differences in tendencies, attitudes, and experiences (OECD, 2013). Chen & Volpe (2002) found that men and women have different attitudes to financial matters, and that women have less interest in finance.

There have been some evidences indicating that men have more confidence in their financial skills than females (Chen & Volpe, 2002; Fonseca et al., 2012; Lusardi & Mitchell, 2011; Zissimopoulos, Karney & Rauer, 2008). Some studies focusing on gender related dissimilarities in investment patterns suggested that female investors were likely to have less confidence in their investment decisions than male investors under similar circumstances (Agnew Harrison, 2015), even when there was no apparent variance in actual knowledge (Goldsmith & Goldsmith, 1997). Females seemed more likely to indicate that they do not know how to solve a financial knowledge test rather than attempt to answer it (Bengtsson et al., 2005; (OECD, 2013). However, some studies found that there was no difference in men and women's level of confidence regarding financial decision making (Berggren & Gonzalez, 2010).

Based on previous evidence on gender differences in risk attitudes, financial literacy and confidence in financial matters, there have been some implications for men and women's investment patterns and financial outcomes. For examples, some researchers found that the lower levels of knowledge, lower confidence and risk adverse mindsets can lead women investors to less profitable investments, because of their higher dependence on more secured yet low return saving products and less reliance on high return investments (Brokešová, 2013; Graham et al., 2002). However, in some specific settings, women's lower confidence can be beneficial for decision (Willows & West, 2015).

To recapitulate, the overview picture developed from the various research on the relationship between gender and financial behaviors emphasizes several factors that differentially influence men and women's financial behavior patterns, and especially suggest that there are benefits and drawbacks related to the fact that women invest their financial resources more conservatively. Nonetheless, there have been studies which do not find these expected gender differences (Bliss & Potter, 2002; Powell & Ansic, 1997; Willows & West, 2015). The implications of these differences on overall financial portfolio performance are inadequate and have yet to be explored.

The ability to manage liability efficiently (or inefficiently) directly influences the risk of accepting high levels of liability and the total wealth possessed by an individual. Furthermore, the amount of liability or debt which an individual is willing to take for their own or their firm has psychological and social implications. Some empirical evidence indicates that high levels of liability are related to increased levels of psychological



distress (Bridges & Disney, 2010; Brown et al., 2005), and mental disorders (Jenkins et al., 2008). High levels of liability are also in accordance with individuals' lower self-perceptions of their ability to manage their financial situation (Lange & Byrd, 1998), and with higher probabilities of quitting their journeys or tasks (Dwyer et al., 2013). Consequently, to acquire effective approaches to improve financial health, it is necessary to understand how people control their own or their business's multiple liability or debts, and whether women differ from men in this regard.

Regarding financial decisions, there have been some previous studies providing arguments and evidence about the different level of risk aversion between men and women (Smith et al., 2006). For examples, women entrepreneurs tend to run smaller firms and their businesses are mainly operated in the service sector (Hernandez et al., 2015). They are also likely to manage firms which have lower liability levels as they attempt to reduce the costs of bankruptcy and are more reluctant to provide the essential guarantees to obtain a loan. Romani et al. (2012) discussed that lower levels of liability in enterprises managed by women may also be explained by other factors associated with gender such as discrimination on the part of credit suppliers. There was also evidence that the decisions of the banks about loan requests are different for men and women, despite the similarities in terms of liquidity and solvency.

Other studies mentioned that there is gender discrimination not only in the final decision of granting of credit but also within the application and negotiation process, which discourages women during such process (Hernandez et al., 2015). The issue of greater risk aversion among women and discrimination on the part of the supplies of credit may affect not only the level of liability of firms, but also the cost and maturity of liability (Hernandez et al., 2015). Besides, a longer term of maturity of debt in companies managed by women would be a sign of their greater risk aversion.

Until now, nevertheless, there has been very little examination of gender differences in liability or debt management and the few studies addressing this issue found inconclusive results. Some studies show that women make unplanned purchases (Hira & Mugenda, 2000; Hira & Loibl, 2008) and carry more liability than men (Goldsmith & Goldsmith, 2006). Meanwhile, other studies did not find gender differences in individuals' satisfaction with their liability or debt levels (Hira & Mugenda, 2000) and some studies demonstrated that women tend to pay more attention to financial budget (Henry et al., 2001). Therefore, the role of gender in liability management is vague and calls for further research (Novailitis et al., 2006).

## **Hypotheses**

### **For the Gender Variable**

Rand et al. (2015), in a report on the activities of Vietnamese SMEs in 2015, mentioned that women are less likely to take risks than men. This statement is similar to the views of previous researchers such as Croson and Gneezy (2009); Charness and Gneezy (2012); the OECD (2013); Zinkhan & Karande (1991); Zissimopoulos et al. (2008); Chen & Volpe (2002); Agnew & Harrison (2015); and Agnew et al. (2003).

In the process of operating an enterprise, managers and business owners base themselves on their abilities and the socio-economic environment, applying trade-off theory (Modigliani and Miller, 1963; Kraus and Litzernberger, 1973) and pecking order theory



(Myers & Majluf, 1984; Myers, 1984) in using internal financing and debt to achieve better results.

Based on that, the author hypothesized the following:

*H1: Male entrepreneurs in small and medium-sized enterprises in Vietnam use a higher debt ratio than female entrepreneurs.*

### **For the Control Variables.**

In addition to the main factor that the gender of the business owner affects the debt ratio, other factors related to the individual entrepreneur and environmental factors inside and outside the business all need to be considered.

Regarding education, according to Chen & Volpe (2002), gender differences in financial literacy are related to education and experience. According to their research, factors such as years of education had a significant impact on knowledge. Specifically, business majors know more than non-business majors, and their knowledge increases with the year of study. Human capital generally increases the quality and consistency of assigned work (Becker, 1964; Mincer, 1974), and is beneficial to the acquisition of external financing (Bruederl et al., 1992; Parker & Van Praag, 2006). While Vos et al (2007), using UK and US data, find that younger and less educated SME owners are more actively seeking external financing, while older SME owners and more educated are less likely to seek outside funding sources.

For firm size, Robb et al. (2010) based on company survey data and found that debt increases as firms grow. According to Oakey (1984), while in the early stages of development, many SMEs have been forced to prepare and seek out investment capital from outside, especially growth-oriented companies. Cassar (2004) concludes that the "larger" small businesses are, the more they rely on long-term debt and external financing, including bank loans. This is consistent with Storey (1994), who found that in the case of SMEs, personal savings of the owner-manager as a source of capital during the start-up phase are more important than with outside finance. Firm size can affect the availability of financial resources to the business. As firms grow, their ability to access banks for loans increases (Petersen and Rajan, 1994). While younger firms are often characterized by ambiguity (Berger & Udell, 1998) due to the lack of an established track record, this can lead to banks and institutions. Other financial institutions are reluctant to lend to these companies.

For corporate age, Gregory et al. (2005) argue that older firms should be less dependent on external funding sources than younger firms. They attribute this to the fact that older companies have more opportunities to accumulate retained earnings than younger firms, so there are more internal funds available to fund their operations. Quarthey (2003) concluded that firm age has a significant impact on access to external finance. Fatoki and Asah (2011) find that SMEs established for more than 5 years have a much higher chance of successfully applying for credit than SMEs established less than five years old.

Regarding international cooperation, according to Maes et al. (2019), exporting firms have significantly higher financial leverage than comparable non-exporting firms, stemming from the use of greater use of short-term financial liabilities in exporting enterprises, because of the challenges and opportunities associated with political risks, exchange

rates, and the cultural and geographical distance between domestic and foreign markets. Export destinations. According to the World Trade Organization (WTO), access to financial resources to support exports is a top concern for SMEs because in addition to the one-time upfront cost (for example, the costs associated with regarding compliance with foreign market regulations and prepared market research). Exporting requires substantial ongoing investment in working capital, as exporting significantly extends the cash conversion cycle of the business (e.g., longer shipping times and associated administrative burdens. to international transactions) (WTO, 2016).

Based on the above points of view, the paper makes the following hypotheses:

*H2: Firm size has an impact on the debt ratio of SMEs in Vietnam.*

*H3: Firm age has an impact on the debt ratio of SMEs in Vietnam.*

*H4: The level of education and training of entrepreneurs has an impact on the debt ratio of SMEs in Vietnam.*

*H5: Export factor has an impact on the debt ratio of SMEs in Vietnam.*

### 3 Research Methodology

#### Data.

The study uses data set by the Central Institute for Economic Management (CIEM), the Institute of Labor and Social Sciences (ILSSA), the Institute of World Development Economics of the United Nations University (UNU-WIDER) and the Faculty of Economics (DOE) of the University of Copenhagen jointly investigated the research. The data sample includes more than 2,500 small and medium enterprises in Vietnam from 2007 to 2015.

The topic uses Bayesian method to estimate the coefficients based on the outstanding advantages of Bayes. Bayesian techniques take advantage of all available information, providing intuitive inferences to solve complex problems based on the nature of probabilities and parameters, suitable for decision making. What makes decisions difficult is the uncertainty about the consequences of a decision due to a lack of knowledge about some relevant facts or parameters. The Bayesian method can quantify that level of uncertainty based on reasonable evidence (O'Hagan, 1994). Bayes is particularly valuable in small sample studies, due to the high degree of uncertainty due to sampling error (Gelman, 2009). On the other hand, the quantitative variables in the research model are expressed in logarithmic form. The advantage of the variable being estimated in logarithmic form is to obtain a higher fit and make better use of the data. According to Delmar (1997), using logarithmic variables to correct for the skewed distribution and thus meet the assumption of the normal distribution of the residuals produces unnecessary outliers.

Based on the hypotheses, the variables included in the research model include:

#### Dependent variable.

InDR: debt ratio from 2007 to 2015.

Debt Ratio = Total Liabilities/Total Assets

#### Explanatory variable.

Gen: business gender; Gen = 1, male; Gen = 0, female.

### Control variables.

lnFS: labor force (full time) from 2007 to 2015 (Firm size, representing the size of the business).

lnFA: number of years of establishment (Firm age, age of business).

Ex: Ex is a binary variable, representing the export factor.

Ex = 1, there are exports; Ex = 0, do not export.

D1, D2 and D3 are binary variables, representing the educational level of the entrepreneur:

D1 = 1, if undergraduate and graduate

D2 = 1, if college/intermediate

D3 = 1, if trained occupation without a degree

D1 = D2 = D3 = 0, if there is no profession.

Following previous studies such as those of Hernandez-Nicolas et al. (2015), Gau et al. (2005), and Nyamita et al. (2014), the study proposes a research model as follows:

### Model.

#### Research Results.

In order to select the appropriate a priori information for a large sample size, the sensitivity through five simulations of normally distributed a priori information as follows will be analyzed (Table 1):

The most suitable simulation will be selected based on comparing the estimated Bayesian models with the criteria Log BF, Log (ML), P (M/y) and DIC. In which, the selected criterion will be the largest mean for Log BF, Log (ML), P (M/y) and the smallest for DIC.

Based on the results of Table 2, Model 1 is selected. The estimated values are shown in Table 3.

According to Gelman and Rubin (1992), Brooks and Gelman (1998), diagnostic Rc values greater than 1.2 for any model parameter are considered non-convergent. In practice,  $Rc < 1.1$  is often used to declare convergence. Therefore, the Table 3 above has a Max Gelman-Rubin Rc value  $< 1.1$  indicating that the MCMC convergence is acceptable for Bayesian analysis.

The author used the histogram to consider the degree of autocorrelation, normal distribution and stability. The resulting histograms show low autocorrelation while the trace plots show good association. Normal distributions can be plotted from density histograms and frequency distribution histograms. Therefore, the MCMC convergence of lnDR can be concluded.

The certainty test was also performed and the results show that the following estimates are not significantly different in terms of the latter mean. MCSE and confidence intervals when baseline normal values for all parameters are adjusted from  $-0.5$  to  $0.5$

**Table 1.** Likelihood model

	$\ln DR \sim N(\mu, \delta)$
Prior distributions:	
Simulation 1	$\alpha_i \sim N(0, 1)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 2	$\alpha_i \sim N(0, 10)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha_i \sim N(0, 100)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 4	$\alpha_i \sim N(0, 1000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 5	$\alpha_i \sim N(0, 10000)$ $\delta^2 \sim \text{Invgamma}(0.01, 0.01)$
$i = 1, 2, 3, 4, 5$	

**Table 2.** Bayesian factor test and model test

	Chan	lnDR			
		Avg DIC	Avg log (ML)	Avg log (BF)	P(M/y)
Simulation 1	3	2.19e + 04	-1.10e + 04	●	<b>0.9971</b>
Simulation 2	3	2.19e + 04	-1.10e + 04	-5.8499	0.0029
Simulation 3	3	2.19e + 04	-1.10e + 04	-14.6893	0.0000
Simulation 4	3	2.19e + 04	-1.10e + 04	-23.8076	0.0000
Simulation 5	3	2.19e + 04	-1.10e + 04	-33.0507	0.0000

Source: Author’s calculation from 2007–2015 SME dataset

with an interval of 0.1. Therefore, it can be said that the Bayesian inference result is reasonable and the estimated model is stable.

See the estimated results from Table 3, in the column representing the mean value, the coefficient of the variable Gene has a positive sign, indicating that male entrepreneurs use a higher debt coefficient than female entrepreneurs, exactly as the hypothesis stated. This result is consistent with the views of most researchers such as Croson and Gneezy (2009), Charness and Gneezy (2012), OECD (2013), Zinkhan & Karande (1991), Zissimopoulos et al. (2008), Chen & Volpe (2002), Agnew & Harrison (2015), Agnew et al. (2003), Barber & Odean (2001), Webster & Ellis (1996) and Rand et al.’s results in Vietnam (2015). The results show that in line with Vietnamese women’s character, like (Harvie and Vo 2009), female entrepreneurs often face more challenges than male entrepreneurs. This knowledge relates to access to finance and education.

**Table 3.** Bayesian simulation results

lnDR						
	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
Gen	.031261	.0392258	.000227	.0313524	-.0455716	.1077743
lnFS	.1894067	.0180949	.000104	.1892843	.1538325	.2246954
lnFA	-.1970239	.0287127	.000166	-.1972312	-.2527847	-.1405268
Ex	.3113661	.0735758	.000427	.3110408	.1679402	.4577568
D1	.2081821	.05981	.000345	.208191	.0910168	.3250688
D2	.1077173	.0552117	.000319	.1074011	.0000629	.2163648
D3	.0013647	.0536821	.00031	.001158	-.1028471	.1070024
_cons	-2.694322	.0932458	.000538	-2.695328	-2.875762	-2.510935
var	2.141458	.0389593	.000227	2.140993	2.067034	2.219466

Number of obs = 6,094

Avgacceptance rate = 1

Avgefficiency: min = .9825

MaxGelman-Rubin Rc = 1

Source: Author's calculation from 2007–2015 SME dataset

For the group of control variables, the results from the estimation table show that firm size has a positive effect on the debt ratio, the results are supported by the point of view of Robb et al. (2010), Oakey (1984), Cassar (2004), Storey (1994), (Petersen and Rajan, 1994) and (Berger & Udell, 1998). This result is also consistent with the fact that in Vietnam, small businesses face many difficulties in accessing loans. There are 47% of surveyed enterprises having difficulty in accessing capital, VCCI (2021).

In contrast to firm size, firm age has a negative effect on debt ratio. This result differs from the view of Fatoki and Asah (2011) but coincides with the view of Gregory et al. (2005). According to the trade-off theory and the pecking order theory, able firms can choose between using equity and debt in order to obtain a cost and profit advantage. Suitable for small and medium enterprises in Vietnam.

The export factor has a positive effect on the debt ratio, which coincides with the point of view of Maes et al (2019). As mentioned in the previous content, according to WTO (2016), the operating costs of exporting enterprises are high, it is necessary to ensure flexible response costs due to many risks and inadequacies. According to research results in Vietnam in 2015, exporting enterprises have informal expenditures that are almost double that of non-exporting enterprises (Rand, 2015). Therefore, export enterprises need to strongly support loans, and the research results are appropriate.

The results of business education and training have a positive effect on the debt ratio, consistent with the views of the authors Chen & Volpe (2002), Becker (1964), Mincer (1974), Bruederl et al. (1992), Parker & Van Praag (2006). In fact, in Vietnam, this result is relevant because many small businesses use informal loans, thus not providing enough

financial data, on the other hand, the level of education and training can also limit restrict access to loans from banks and financial institutions.

## 4 Conclusion

The purpose of the study is to study the effect of the gender difference of entrepreneurs on the debt coefficient of small and medium enterprises in Vietnam through Bayesian estimation method. Research results show that the debt ratio in male-owned enterprises is higher than the debt ratio in female-owned enterprises. In addition, the research results also show that enterprise size, export factors, and education and training level of entrepreneurs have a positive impact on the debt ratio, while the age of the enterprise has a negative effect to the debt ratio.

This can be the necessary reference result for managers, business owners, exporters, educators, banks and policy makers.

## References

- Agnew, J., Balduzzi, P., Sundén, A.: Portfolio choice and trading in a large 401 (k) plan. *Am. Econ. Rev.* 193–215 (2003)
- Agnew, S., Harrison, N.: Financial literacy and student attitudes to debt: A cross national study examining the influence of gender on personal finance concepts. *J. Retail. Consum. Serv.* **25**, 122–129 (2015). <https://doi.org/10.1016/j.jretconser.2015.04.006>
- Arano, K., Parker, C., Terry, R.: Gender-based risk aversion and retirement asset allocation. *Econ. Inq.* **48**(1), 147–155 (2010). <https://doi.org/10.1111/j.1465-7295.2008.00201.x>
- Barber, B.M., Odean, T.: Boys will be boys: gender, overconfidence, and common stock investment. *Q. J. Econ.* 261–292 (2001)
- Becker, G.S.: *Human capital*. Columbia University Press, New York (1964)
- Bengtsson, C., Persson, M., Willenhag, P.: Gender and overconfidence. *Econ. Lett.* **86**(2), 199–203 (2005). <https://doi.org/10.1016/j.econlet.2004.07.012>
- Berger, A.N., Udell, G.F.: The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle. *J. Bank. Financ.* **22**, 613–673 (1998). [https://doi.org/10.1016/S0378-4266\(98\)00038-7](https://doi.org/10.1016/S0378-4266(98)00038-7)
- Berggren, J., Gonzalez, R.: Gender difference in financial decision making- A quantitative study of risk aversion and overconfidence between the genders, Umeå University (2010). <https://www.diva-portal.org/smash/get/diva2:324378/FULLTEXT01.pdf>
- Bliss, R.T., Potter, M.E.: Mutual fund managers: does gender matter? *J. Bus. Econ. Stud.* **8**(1), 1 (2002). [https://www.researchgate.net/publication/288943204\\_Mutual\\_fund\\_managers\\_Does\\_gender\\_matter](https://www.researchgate.net/publication/288943204_Mutual_fund_managers_Does_gender_matter)
- Bokešová, Z.: Gender differences in financial decisions. In: *GV-CONF 2013: Proceedings in Global Virtual Conference*, pp. 119–122 April (2013)
- Boyd, B.K.: Board control and ceo compensation. *Strateg. Manag. J.* **15**(5), 335–344 (1994). <https://doi.org/10.1002/smj.4250150502>
- Bridges, S., Disney, R.: Debt and depression. *J. Health Econ.* **29**(3), 388–403 (2010). <https://doi.org/10.1016/j.jhealeco.2010.02.003>
- Brook, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**(4), 434–455 (1998). <https://doi.org/10.1080/10618600.1998.10474787>
- Brown, S., Taylor, K., Price, S.W.: Debt and distress: Evaluating the psychological cost of credit. *J. Econ. Psychol.* **26**(5), 642–663 (2005). <https://doi.org/10.1016/j.joep.2005.01.002>

- Bruederl, J., Preissendoerfer, P., Ziegler, R.: Survival chances of newly founded business organizations. *Am. Sociol. Rev.* **57**, 227–242 (1992). <https://doi.org/10.2307/2096207>
- Brush, C.: Research on women business owners: Past trends, a new perspective and future directions. *Entrepreneurship Theory Pract.* **16**, 5–26 (1992). <https://doi.org/10.1177/104225879201600401>
- Byrnes, J.P., Miller, D.C., Schafer, W.D.: Gender differences in risk taking: A meta-analysis. *Psychol. Bull.* **125**(3), 367 (1999). <https://doi.org/10.1037/0033-2909.125.3.367>
- Cassar, G.: The financing of business start-ups. *J. Bus. Ventur.* **19**(2), 261–283 (2004). [https://doi.org/10.1016/S0883-9026\(03\)00029-6](https://doi.org/10.1016/S0883-9026(03)00029-6)
- Charness, G., Gneezy, U.: Strong evidence for gender differences in risk taking. *J. Econ. Behav. Organ.* **83**(1), 50–58 (2012). <https://doi.org/10.1016/j.jebo.2011.06.007>
- Chen, H., Volpe, R.P.: Gender differences in personal financial literacy among college students. *Financ. Serv. Rev.* **11**(3), 289–307 (2002)
- Croson, R., Gneezy, U.: Gender differences in preferences. *J. Econ. Lit.* 448–474 (2009). <https://doi.org/10.1257/jel.47.2.448>
- Davies, E., Lea, S.E.: Student attitudes to student debt. *J. Econ. Psychol.* **16**(4), 663–679 (1995). [https://doi.org/10.1016/0167-4870\(96\)80014-6](https://doi.org/10.1016/0167-4870(96)80014-6)
- Delmar, F.: *Measuring Growth: Methodological Considerations and Empirical Results in Entrepreneurship and SME Research: On its Way to the Next Millennium*. Routledge (1997)
- Donaldson, G.: *Corporate debt capacity: A study of corporate debt policy and the determination of corporate debt capacity*. Division of Research, Harvard Graduate School of Business Administration, 1961 (1961). ISBN 1–58798–034–7
- Dwyer, R.E., Hodson, R., McCloud, L.: Gender, debt, and dropping out of college. *Gen. Soc.* **27**(1), 30–55 (2013). <https://doi.org/10.1177/0891243212464906>
- Eckel, C.C., Grossman, P.J.: Differences in the economic decisions of men and women: Experimental evidence. *Handb. Exp. Econ. Results* **1**, 509–519 (2008)
- Fatoki, O., Asah, F.: The impact of firm and entrepreneurial characteristics on access to debt finance by SMEs in king williams' town, South Africa. *Int. J. Bus. Manag.* **6**(8), 170–179 (2011). <https://doi.org/10.5539/ijbm.v6n8p170>
- Feng, L., Seasholes, M.S.: Individual investors and gender similarities in an emerging stock market. *Pac. Basin Financ. J.* **16**(1), 44–60 (2008). <https://doi.org/10.1016/j.pacfin.2007.04.003>
- Fonseca, R., Mullen, K.J., Zamarró, G., Zissimopoulos, J.: What explains the gender gap in financial literacy? The role of household decision making. *J. Consum. Aff.* **46**(1), 90–106 (2012). <https://doi.org/10.1111/j.1745-6606.2011.01221.x>
- Gaud, P., Jani, E., Hoesli, M., Bender, A.: the capital structure of Swiss companies: An empirical analysis using dynamic panel data. *Eur. Financ. Manag.* **11**(1), 5–69 (2005). <https://doi.org/10.1111/j.1354-7798.2005.00275.x>
- Gelman, A.: Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Stat. Sci.* **24**(2), 176–178 (2009)
- Gelman, A., Rubin, D.B.: Inference form iterative simulation using multiple sequences. *Stat. Sci.* **7**(4), 457–511 (1992). <https://doi.org/10.1214/ss/1177011136>
- Goldsmith, R.E., Goldsmith, E.B.: The effects of investment education on gender differences in financial knowledge. *J. Pers. Financ.* **5**(2), 55–69 (2006)
- Goldsmith, E., Goldsmith, R.E.: Gender differences in perceived and real knowledge of financial investments. *Psychol. Rep.* **80**(1), 236–238 (1997). <https://doi.org/10.2466/pr0.1997.80.1.236>
- Gregory, B.T., Rutherford, M.W., Oswald, S., Gardiner, L.: An empirical investigation of the growth cycle theory of small firm financing. *J. Small Bus. Manage.* **43**(4), 382–392 (2005). <https://doi.org/10.1111/j.1540-627X.2005.00143.x>
- Graham, J.F., Stendardi Jr, E.J., Myers, J.K., Graham, M.J.: Gender differences in investment strategies: an information processing perspective. *Int. J. Bank Mark.* **20**(1), 17–26 (2002)

- Halko, M.L., Kaustia, M., Alanko, E.: The gender effect in risky asset holdings. *J. Econ. Behav. Organ.* **83**(1), 66–81 (2012). <https://doi.org/10.1016/j.jebo.2011.06.011>
- Harvie, C., Vo, A.N.: The changing face of women managers in small and medium sized enterprises in Vietnam. In: Truong, Q., Rowley, C. (eds.) *The Changing Face of Vietnamese Management*, pp. 221–250. Routledge, New York (2009). <https://doi.org/10.4324/9780203868409>
- Hernandez Bark, A.S., Escartín, J., Schuh, S.C., van Dick, R.: Who leads more and why? A mediation model from gender to leadership role occupancy. *J. Bus. Ethics* **139**(3), 473–483 (2015). <https://doi.org/10.1007/s10551-015-2642-0>, <https://www.jstor.org/stable/44164237>
- Hernandez-Nicolas, C.M., Martín-Ugedo, J.F., Mínguez-Vera, A.: the influence of gender on financial decisions: evidence from small start-up firms in Spain. *Bus. Adm. Manag.* **18**(4), 93–107 (2015). <https://doi.org/10.15240/tul/001/2015-4-007>
- Henry, R.A., Weber, J.G., Yarbrough, D.: Money management practices of college students. *Coll. Stud. J.* **35**(2), 244 (2001)
- Hira, T.K., Mugenda, O.: Gender differences in financial perceptions, behaviors and satisfaction. *J. Fin. Plan.-Denver* **13**(2), 86–93 (2000)
- Hira, T.K., Loibl, C.: Gender differences in investment behavior, pp. 253–270. Springer, New York (2008). [https://doi.org/10.1007/978-0-387-75734-6\\_15](https://doi.org/10.1007/978-0-387-75734-6_15)
- Jenkins, R., et al.: Debt, income and mental disorder in the general population. *Psychol. Med.* **38**(10), 1485–1493 (2008). <https://doi.org/10.1017/S0033291707002516>
- Judge, W.Q., Zeithaml, C.: Institutional and strategic choice perspectives on board involvement in the strategic decision process. *Acad. Manag. J.* **35**, 766–794 (1992). <https://doi.org/10.2307/256315>
- Kraus, A., Litzberger, R.H.: A state-preference model of optimal financial leverage. *J. Financ.* **28**(4), 911–922 (1973). <https://doi.org/10.2307/2978343>
- Kunnanatt, J.T., Emiline, M.: Investment strategies and gender: a study of emerging patterns in India. *J. Gen. Stud.* **21**(4), 345–363 (2012). <https://doi.org/10.1080/09589236.2012.661569>
- Lange, C., Byrd, M.: The relationship between perceptions of financial distress and feelings of psychological well-being in New Zealand university students. *Int. J. Adolesc. Youth* **7**(3), 193–209 (1998). <https://doi.org/10.1080/02673843.1998.9747824>
- Lusardi, A., Mitchell, O.S.: Financial literacy around the world: an overview. *J. Pension Econ. Financ.* **10**(04), 497–508 (2011). <https://doi.org/10.1017/S1474747211000448>
- Maes, E., Dewaelheyns, N., Fuss, C., Van Hulle, C.: The impact of exporting on financial debt choices of SMEs. *J. Bus. Res.* **102**, 56–73 (2019). <https://doi.org/10.1016/j.jbusres.2019.05.008>
- Mincer, J.: *Schooling, experience, and earnings*. Columbia University Press, New York (1974)
- Modigliani, F., Miller, M.H.: Corporate income taxes and the cost of capital: A correction. *Am. Econ. Rev.* **53**(3), 433–443 (1963). <https://www.jstor.org/stable/1809167>
- Corporate financing and investment decisions when firms have information that investors do not have. *J. Financ. Econ.* **13**, 187–221 (1984). [https://doi.org/10.1016/0304-405X\(84\)90023-0](https://doi.org/10.1016/0304-405X(84)90023-0)
- Myers, S.C.: The capital structure puzzle. *J. Financ.* **39**(3), 575–591 (1984). <https://doi.org/10.1111/j.1540-6261.1984.tb03646.x>
- Novailitis, J.M., Merwin, M.M., Osberg, T.M., Roehling, P.V., Young, P., Kamas, M.M.: Personality factors, money attitudes, financial knowledge, and credit-card debt in college students I. *J. Appl. Soc. Psychol.* **36**(6), 1395–1413 (2006). <https://doi.org/10.1111/j.0021-9029.2006.00065.x>
- Nyamita, N.O., Garbharran, H.L., Dorasamy, N.: Factors influencing debt financing decisions of corporations – theoretical and empirical literature review. *Probl. Perspect. Manag.* **12**(4), 189 – 202 (2014). <http://hdl.handle.net/10321/1203>
- Oakey, R., P.: Finance and Innovation in British Small Independent Firms. *Int. Journal Manag. Sci.* **12**(2), 113–124 (1984). [https://doi.org/10.1016/0305-0483\(84\)90030-6](https://doi.org/10.1016/0305-0483(84)90030-6)
- OECD (2013). *Financial literacy and inclusion: Results of OECD/INFE survey across countries and by gender*. The Russia Financial Literacy and Education Trust Fund



- O'Hagan, A.: Kendall's advanced theory of statistics, vol. 2B. Bayesian inference. Arnold, London (1994)
- Parker, S.C., Van Praag, M.: Schooling, capital constraints, and entrepreneurial performance: The endogenous triangle. *J. Bus. Econ. Stat.* **24**(4), 416–431 (2006). <https://www.jstor.org/stable/27638893>
- Petersen, M.A., Rajan, R.G.: The benefits of lending relationships: Evidence from small business data. *J. Finance* **49**(1), 3–37 (1994). <https://doi.org/10.1111/j.1540-6261.1994.tb04418.x>
- Powell, M., Ansic, D.: Gender differences in risk behaviour in financial decision-making: an experimental analysis. *J. Econ. Psychol.* **8**(6), 605–628 (1997)
- Quartey, P.: Financing Small and Medium enterprises (SMEs) in Ghana. *J. Afr. Bus.* **4**(1), 37–55 (2003). [https://doi.org/10.1300/J156v04n01\\_03](https://doi.org/10.1300/J156v04n01_03)
- Rand, J., Brandt, K., Sharma, S., Trifkovic, N.: Characteristics of the Vietnamese business environment: evidence from a SME survey in 2015, The Central Institute for Economic Management (CIEM), the Institute of Labor Science and Social Affairs (ILSSA), the United Nations University Institute for World Development Economics (UNU-WIDER), and the Department of Economics (DOE) of the University of Copenhagen(2015)
- Robb, A., Reedy, E., Ballou, J., DesRoches, D., Potter, F., Zhao, Z.: An overview of Kauffman Firm Survey: Results from the 2004–2008 data. The Ewing Marion Kauffman Foundation (2010)
- Romani, G., Atienza, M., Amorós, J.E.: Informal investors in Chile: an exploratory study from a gender perspective. *J. Bus. Econ. Manag.* **13**(1), 111–131 (2012). <https://doi.org/10.3846/1611699.2011.620141>
- Schubert, R., Brown, M., Gysler, M., Brachinger, H.W.: Financial decision-making: are women really more risk-averse? *Am. Econ. Rev.* **89**(2), 381–385 (1999). <https://www.jstor.org/stable/117140>
- Siva, S.: A study on gender difference in investment choice & risk-taking. *IJAR BAE* **1**(1), 01–06 (2012)
- Sladek, R.M., Bond, M.J., Phillips, P.A.: Age and gender differences in preferences for rational and experiential thinking. *Personality Individ. Differ.* **49**(8), 907–911 (2010). <https://doi.org/10.1016/j.paid.2010.07.028>
- Smith, N., Smith, V., Verner, M.: Do women in top management affect firm performance? A panel study of 2,500 Danish firms. *Int. J. Product. Perform. Manag.* **55**, 569–593 (2006). <https://doi.org/10.1108/17410400610702160>
- Soltane, B.: Governance and performance of microfinance institutions in Mediterranean countries. *J. Bus. Econ. Manag.* **10**(1), 31–43 (2009). <https://doi.org/10.3846/1611-1699.2009.10.31-43>
- Storey, D.J.: Understanding the small business sector. 20 years of Entrepreneurship Research, Swedish Entrepreneurship Forum 2014 (1994). ISBN: 91–89301–56–0
- Terjesen, S., Sealy, R., Singh, V.: Women directors on corporate boards: A review and research agenda. *Corp. Gov. Int. Rev.* **17**, 320–337 (2009). <https://doi.org/10.1111/j.1467-8683.2009.00742.x>
- Vietnam Chamber of Commerce and Industry (VCCI). Vietnam's Provincial Competitiveness Index: Assessing the Quality of Economic Governance to Promote Business Development (PCI 2021). Hanoi: Vietnam Chamber of Commerce and Industry Vnexpress (2021). <https://e.vnexpress.net/news/business/data-speaks/vietnamese-women-outnumbered-by-men-in-top-management-roles-report-3727865.html>
- Vos, E., Yeh, S., Carter, S., Tagg, S.: The happy story of small business financing. *J. Bank. Finance* **31**(9), 2648–2672 (2007). <https://doi.org/10.1016/j.jbankfin.2006.09.011>
- Weisbach, M.S.: Outside directors and CEO turnover. *J. Financ. Econ.* **20**, 431–460 (1988). [https://doi.org/10.1016/0304-405X\(88\)90053-0](https://doi.org/10.1016/0304-405X(88)90053-0)

- Willows, G., West, D.: Differential Investment performance in South Africa based on gender and age. *Int. Bus. Econ. Res. J. (IBER)*, **14**(3), 537–560 (2015). <https://doi.org/10.19030/iber.v14i3.9215>
- WTO (2016). Trade finance and SMEs: Bridging the gaps in provision, pp. 1–44 [https://www.wto.org/english/res\\_e/booksp\\_e/tradefinsme\\_e.pdf](https://www.wto.org/english/res_e/booksp_e/tradefinsme_e.pdf)
- Zinkhan, G.M., Karande, K.W.: Cultural and gender differences in risk-taking behavior among American and Spanish decision makers. *J. Soc. Psychol.* **131**(5), 741–742 (1991). <https://doi.org/10.1080/00224545.1991.9924657>
- Zissimopoulos, J.M., Karney, B., Rauer, A.: Marital histories and economic well-being. Working Paper, WP, 180 (2008). <https://deepblue.lib.umich.edu/handle/2027.42/61806>



# Contagion Effects Among Commodity Markets and Securities Markets During the Conflict Between Russia and Ukraine: The Dynamic Conditional Correlation Approach

Sunisa Phaimekha and Worrawat Saijai<sup>(✉)</sup>

Center of Excellence in Econometrics, Faculty of Economics,  
Chiang Mai University, Chiang Mai 50200, Thailand  
[worrawat.s@cmu.ac.th](mailto:worrawat.s@cmu.ac.th)

**Abstract.** The continuation of the Russia-Ukraine war has led to an interest in examining the impacts of this war on the volatilities of various financial markets from February 2022 to May 2022 by using pre-war and wartime data covering the period from January 2010 to May 2022. The commodity and securities markets are considered, and the dynamic correlation between the volatilities of different financial markets is measured using the dynamic conditional correlation (DCC) based on the multivariate GARCH model. The DCC allows analysis of the extent of the impact. Results indicate that all return series display persistently high volatility at values greater than 0.80. Comparing the extent of pre-war and wartime impacts, following the start of the war there appears to be an increase in the conditional correlations but a decrease in the correlation between the volatilities of several financial market pairs, indicating that the impact between these markets exists. Moreover, some assets can serve as a safe haven for other assets.

**Keywords:** GARCH · DCC-GARCH · Russia-Ukraine · Safe haven

## 1 Introduction

Commodities are goods that are frequently used as raw materials in manufacturing and are known as products that must be used in large quantities continuously. Therefore, when trading, it is necessary to have futures contracts to hedge the risk as prices may change at any time. Since commodities have the same global standard, the direction of price changes is based on global supply and demand. In addition, commodities are assets for investment because their prices are generally positively correlated with inflation. Investing in commodities has the advantage of being able to adjust the value of an investment according to inflation, and their prices tend to react quickly to immediate events.

Decades of certainty are no longer valid following the COVID-19 outbreak, and then one of the greatest conflicts in the world came up. The conflict between Russia and Ukraine led to the war between the two countries. It not only endangers human life and property; it also raises the risks to the global economy and trade. The high degree of uncertainty in the situation can affect us in many ways. One of them is the commodity price. Oil and gas prices will increase due to the Western countries' sanctions with the purpose of increasing economic pressure on Russia to stop the war (Huther 2022), and the prices of related goods will also soar, e.g., wheat, soybeans, corn, etc. The increasing commodity prices lead to high inflation worldwide, directly affecting people's spending decisions. The globe is getting more complicated, risks and uncertainties increase, challenges become more complex, and the countries require faster and more consistent action to deal with the upcoming economic recession. Against this background, it is imperative to describe what can and must be learned from the current crisis (Fischedick 2022). This requires a technique that is based on a dynamic concept to provide a better understanding of the current situation.

Bitcoin is similar to commodities, and the analysis of Bitcoin prices can be generally learned from the analysis of resource commodities (Gronwald 2019). One study said that Bitcoin is not money but rather a digital commodity with value but no value-added because both the production of and the speculation with Bitcoin draw from the existing global pool of value-added (Rotta 2022). So, it can be said that Bitcoin is a commodity. Moreover, Bitcoin was used instead of the ruble during the Russia-Ukraine war in some transactions. One of the reasons is that the European Union banned some Russian banks from using the Society for Worldwide Interbank Financial Telecommunication (SWIFT) to put pressure on the Russian government during this war. And also, some Ukrainians are turning to Bitcoin as an alternative to Ukrainian financial institutions. In a scenario where governments are in chaos, relying on traditional banks is difficult, and there is fear of surveillance. So, a relatively anonymous system where no government is involved is appealing. Then, it is interesting to study Bitcoin as a commodity in this paper.

Volatility has no definitive outcome. This is because each fluctuation depends on the data and the computed model. In this paper, we focus on calculating the volatility errors in the relationship between the dominant securities in the financial markets and the commodities during the war between Russia and Ukraine. Since the volatilities are dynamic, the generalized autoregressive heteroscedasticity (GARCH) model is employed to describe them. The dynamic conditional correlation (DCC) model is used to describe the fluctuation of correlation.

In the time of conflict between Russia and Ukraine, the correlation between different financial and commodity markets' volatilities described in this paper can be applied as one of the academic insights that provide facts about the correlation and how the co-movement might change in these markets during the historically volatile period. This will be beneficial to policymakers as the existence of this correlation can be confirmed.

The rest of this paper is structured as follows: Sect. 2 presents the methodology, which mainly deals with the GARCH model. Section 3 reports the data and preliminary analysis. Section 4 reports the empirical findings. And Sect. 5 gives the conclusions.

## 2 Methodology

The DCC-GARCH model consists of two parts: the volatility model (GARCH process) and the dynamic conditional correlation (DCC) model. This study employs three GARCH-like models, including GARCH, GJR-GARCH, and CS-GARCH, to capture financial market volatility. These models are helpful in studying volatility, assuming that variability is dynamic rather than stable. These models have several advantages because they provide reliable results by eliminating the autocorrelation and heterogeneity common in time series data. For greater flexibility, this section uses the DCC model to explain the correlation with a better dynamic concept when the correlation changes over time.

### 2.1 GARCH Models

The standard model comprises mean and variance equations. The first equation is the mean equation written as

$$r_{i,t} = \mu_i + \varepsilon_{i,t}, \quad i = 1, 2, \dots, n, \quad (1)$$

and

$$\varepsilon_{i,t} = \sqrt{h_{i,t}} z_{i,t}, \quad (2)$$

The stock  $i$ 's return at time  $t$  is represented by  $r_{i,t}$ ,  $\mu_i$  is the return's average which is a constant term, and the residual term is  $\varepsilon_{i,t}$ .  $z_{i,t}$  is a standardized residual. Both  $\varepsilon_{i,t}$  and  $z_{i,t}$  are Independent and identically distributed (*i.i.d.*) random variables, with the property of *i.i.d.*,  $E(z_{i,t}) = 0$  and  $E(z_t z_t^T) = I$ , where  $z_t$  is a vector consisting of the values  $z_{i,t}$  for different  $i$ .

Under the GARCH(1,1) process, the below equation describes conditional variance  $h_{i,t}$  of return at time  $t$  is

$$h_{i,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1}, \quad (3)$$

In the variance equation, all parameters, including  $\omega_i, \alpha_i, \beta_i > 0$  have to be positive.  $\alpha$  captures ARCH effect on volatility. The GARCH effect is presented by  $\beta$ , showing the stackability of  $h_{i,t}$  from its own past value. In the other words, ARCH and GARCH effects reveal the existence and persistence of the volatility, respectively. To extend more, ARCH shows the persistence in residual and GARCH shows the persistence in conditional variance.

**2.2 GJosten-Jagannathan-Runkle-GARCH (GJR-GARCH)**

The GJR-GARCH is an alternative GARCH model providing leverage term  $\gamma_i$ . The leverage term helps capture the volatility’s bad news and good news. When the error  $\varepsilon_{i,t-1}$  is negative,  $I_{i,t-1}$  equals to one, otherwise, it is zero.

$$h_{i,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1} + \gamma_i I_{i,t-1} h_{i,t-1}, \tag{4}$$

where

$$I_{i,t-1} = \begin{cases} 1 & \text{if } \varepsilon_{i,t-1} < 0 \\ 0 & \text{if } \varepsilon_{i,t-1} \geq 0 \end{cases}, \tag{5}$$

**2.3 Component GARCH (COGARCH)**

The Component GARCH model provides a leverage term using  $q_{i,t}$ , which is an additional ARCH(1) quantity. The COGARCH(1,1) can be shown as follows

$$h_{i,t} = q_{i,t} + \alpha_i (\varepsilon_{i,t-1}^2 - q_{i,t-1}) + \beta_i (h_{i,t-1} - q_{i,t-1}), \tag{6}$$

$$q_{i,t} = v + \delta q_{i,t-1} + \phi (\varepsilon_{i,t-1}^2 - h_{i,t-1}), \tag{7}$$

where  $v, \delta$  and  $\phi$  are the estimated parameters.

**2.4 Dynamic Conditional Correlation (DCC)**

The DCC model is more realistic than the static correlation methods, such as Pearson and Spearman correlations which can not measure the time-varying correlation between the random variables. Engle (2002) suggested that the correlation could not be constant and there might exist structural change in the correlation series. In the other words, the behavior of the variables has been changing over time, thus, the interaction between variables could be also changed. To measure this time-varying correlation, Engle (2002) suggested the DCC process predicts the correlation at time  $t$ . In the DCC model, the correlation matrix is  $R_t$ , and the covariance,  $H_t$ . The covariance is affected by its own past at  $t - 1$ , which is called the conditional covariance,  $H_t = E[r_t r_t' | \Psi_{t-1}]$  where  $\Psi_{t-1}$  is a set of past innovations (Engle 2002), or it is the past value of return in our case. The covariance matrix can be computed by using

$$H_t = D_t R_t D_t, \tag{8}$$

where  $D_t$  is a diagonal matrix of the time-varying conditional standard deviation of the error  $\varepsilon_{1,t}$  under the GARCH(1,1) process. The matrix of  $D_t$  with the  $n \times n$  dimension can be shown as follows

$$D_t = \begin{bmatrix} \sqrt{h_{1,t}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{h_{2,t}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{h_{3,t}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{h_{n,t}} \end{bmatrix}, \tag{9}$$

The correlation matrix,  $R_t$ , is a symmetric positive semi-definite matrix, showing a correlation between asset  $i$  and  $j$ ,  $\rho_{ij}, i \neq j$

$$R_t = \begin{bmatrix} 1 & \rho_{12,t} & \cdots & \rho_{1n,t} \\ \rho_{21,t} & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1,t} & \rho_{n2,t} & \cdots & 1 \end{bmatrix}, \tag{10}$$

The parameter set ( $\Theta$ ) can be computed using the DCC log-likelihood function as

$$L(\Theta) = -\frac{1}{2} \sum_{t=1}^T (n \log(2\pi) + \log |D_t| + \log |R_t| + \varepsilon_t' R_t^{-1} \varepsilon_t). \tag{11}$$

where  $|D_t|$  and  $|R_t|$  are the determinant of the matrix D and R respectively. And the Maximum likelihood estimation is employed to find the best-explained parameters.

### 3 Data

We investigate the dynamic correlation between stock returns of the Dow Jones Industrial Average (DJI), the 10-Year Treasury Yield of the USA (TNX) of NY Mercantile, Wheat (WHT), Corn (CRN), Soybean (SOY), Sugar (SUG), NYMEX West Texas Intermediate crude oil (WTI), natural gas (NG), the gold price (GOLD) of COMEX, silver (SIL), copper (COP), and bitcoin (BTC). The data is collected by Yahoo Finance.com. The sample covers the period from January 4, 2010, to May 10, 2022, with 1845 observations. We include this range because we want to explore the dynamic correlations between commodity and securities markets before and during the war between Russia and Ukraine. This will give us a clearer understanding of the implications of this war. Table 1 and Table 2 show negative skewness and high kurtosis. The results show that the data are not normally distributed after finding the minimum Bayesian factor (MBF). Moreover, the extended Dickey-Fuller test (unit root) shows that all returns are stationary. Our goal is to examine the degree of contagion in financial markets.

**Table 1.** Descriptive statistics of returns (1)

	DJI	TNX	WHT	CRN	SOY	SUG
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Median	0.001	0.000	-0.000	0.001	0.001	0.000
Maximum	0.108	0.017	0.197	0.077	0.064	0.108
Minimum	-0.138	-0.015	-0.113	-0.191	-0.086	-0.078
Std. Dev.	0.012	0.003	0.019	0.016	0.012	0.019
Skewness	-1.054	-0.110	0.610	-0.930	-0.223	0.308
Kurtosis	23.548	2.295	7.673	13.176	3.537	1.895
Jarque-Bera	43,075*	410*	4,654*	13,648*	980*	306*
Unit root test (ADF)	-29.554*	-31.769*	-28.679*	-30.432*	-29.982*	-30.046*

Note: “\*” denotes a strong rejection of the null hypothesis, according to Minimum Bayes Factor (MBF) (see, Maneejuk and Yamaka 2021)

**Table 2.** Descriptive statistics of returns (2)

	WTI	NG	GOLD	SIL	COP	BTC
Mean	0.001	0.001	0.000	0.000	0.000	0.003
Median	0.002	0.000	0.000	0.000	0.000	0.002
Maximum	0.320	0.382	0.058	0.089	0.064	0.225
Minimum	-0.282	-0.301	-0.051	-0.124	-0.069	-0.465
Std. Dev.	0.031	0.035	0.009	0.018	0.014	0.047
Skewness	0.123	0.436	-0.120	-0.732	-0.077	-0.655
Kurtosis	21.988	12.931	4.472	7.074	1.408	8.785
Jarque-Bera	37,262*	12,946*	1,547*	4,023*	155*	6,082*
Unit root test (ADF)	-30.582*	-32.695*	-30.780*	-29.076*	-31.304*	-30.240*

Note: “\*” denotes a strong rejection of the null hypothesis, according to Minimum Bayes Factor (MBF) (see, Maneejuk and Yamaka 2021)

## 4 Estimation Results

Before figuring out what the results mean for volatility and correlation, we look at different DCC-GARCH models with four different distributional assumptions: Student’s t, Student’s skewed, normal, and normally skewed. Table 3 shows that the normally distributed GARCH (1,1) has the lowest value of the Bayesian Information Criterion (BIC) and is thus the best model for explaining the properties of the data.

**Table 3.** BIC estimates for multivariate GARCH (1,1) under four distributions

	Std	Sstd	Norm	Snorm
GJR-GARCH	-70.519	-70.470	-70.565	-70.517
GARCH	-70.524	-70.478	-70.583	-70.533
CSGARCH	-70.475	-70.422	-70.537	-70.484



In Table 4, we show only the best estimation results of the GARCH (1,1) obtained from the previous step. According to Table 4, a high level of volatility persistence is measured by  $\alpha_i + \beta_i$ , in all markets as the values along our sample period are higher than 0.8. Table 4 also shows that copper is the least volatile asset (0.8507). On the other hand, natural gas is the most volatile (0.9990). In addition, the results also provide the DCC parameters ( $a$  and  $b$ ). A dynamic conditional correlation can be interpreted as persistent if the sum between these parameters is close to 1.

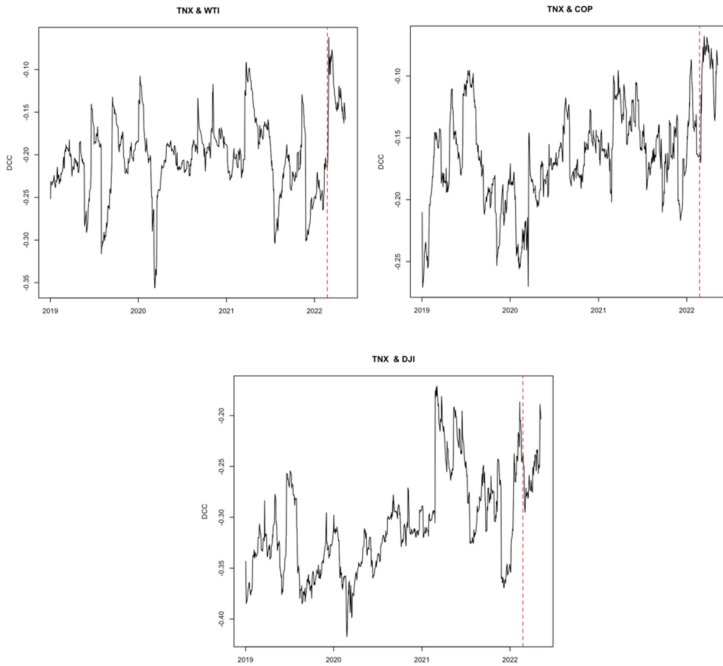
**Table 4.** Results of the DCC-GARCH model

		DJI	TNX	WHT	CRN	SOY	SUG	WTI	NG	GOLD	SIL	COP	BTC
Mean Eq.	$\mu_i$	0.001*	-0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	-0.000	0.000	0.003*
GARCH	$\omega_i$	0.000	0.000	0.000*	0.000	0.000	0.000	0.000*	0.000	0.000	0.000	0.000	0.000*
	$\alpha_i$	0.220*	0.058*	0.092*	0.077	0.070*	0.050	0.139*	0.100*	0.023	0.033*	0.073	0.144*
	$\beta_i$	0.744*	0.919*	0.823*	0.903*	0.909*	0.874*	0.835*	0.899*	0.970*	0.961*	0.777*	0.801*
	$a$	0.009*											
	$b$	0.944*											

Note: “\*\*” denotes a strong rejection of the null hypothesis, according to Minimum Bayes Factor (MBF) (see, Maneejuk and Yamaka 2021)

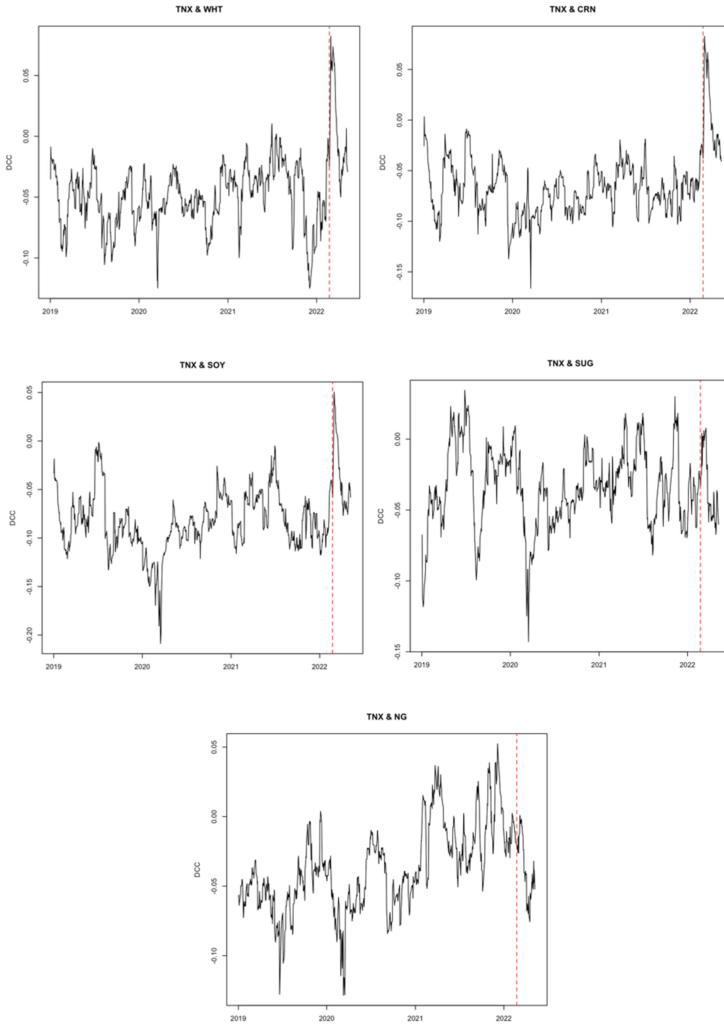
The result shows that conditional volatility increased after the war between Russia and Ukraine arose. Since then, it has shown a decline in many markets. Except for NG, which suddenly rises and then falls. Before the war, NG, WHT, and WTI were the top three markets showing a great reaction to the war, with highs of 0.025, 0.005, and 0.004, respectively. The degree of correlation between financial markets was found to mostly increase during this war, linked to the potential impact of the war.

According to the result, TNX is always an alternative investment that reduces volatility or is a safe haven asset for WTI, COP, and DJI, as a negative correlation is shown in Fig. 1. However, TNX only served as a safe-haven asset for WHT, CRN, SOY, and SUG prior to the war. But when the war began, TNX was no longer a safe haven for WHT, CRN, SOY, and SUG. This differs from NG in that TNX gradually became a safe-haven asset for it after the war began, as shown by the correlation in Fig. 2.



**Fig. 1.** Dynamic conditional correlation: TNX - WTI, COP, and DJI. The vertical red dashed line is the date the Russian-Ukrainian conflict started (February 24, 2022).

The United States of America is sensitive to the war between Russia and Ukraine, as the correlations with other market prices have obviously increased showing that the United States is not the best choice to invest in during a war if the investors want to take the lower risk. However, TNX seems to be the best place to park money with the lowest volatility. Moreover, COP and BTC have not increased in correlation compared to other stocks. This evidence allows us to conclude that financial markets are affected by the war. In addition, TNX negatively correlates with WTI, COP, DJI, NG, and SUG during the war, confirming that TNX is the safe haven for WTI, COP, DJI, NG, and SUG. Investors in these markets are encouraged to invest in the TNX market to mitigate the risk to their portfolios.



**Fig. 2.** Dynamic conditional correlation: TNX - WHT, CRN, SOY, SUG, and NG. The vertical red dashed line is the date the Russian-Ukrainian conflict started (February 24, 2022).

## 5 Conclusion

This paper aims to examine the correlation among financial markets by detecting the impact of the war between Russia and Ukraine. The correlation between different investments changes with time, so the relationship between the assets should not be fixed; thus, DCC-GARCH-type models are used to model the correlation. The data employed in this study was collected from January 4, 2010, to May 10, 2022, encompassing 1845 observations. The market data include

Dow Jones Industrial Average (DJI), 10-Year Treasury Yield of USA (TNX) of NY Mercantile, Wheat (WHT), Corn (CRN), Soybean (SOY), Sugar (SUG), NYMEX West Texas Intermediate crude oil (WTI), Natural gas (NG), Gold Price (Gold) of COMEX, Silver (SIL), Copper (COP) and Bitcoin (BTC). The model selection result indicates that the multivariate S-GARCH model outperforms other multivariate GARCH models. Furthermore, the result of the multivariate DCC-GARCH shows that some of the assets performed moderately after the war between Russia and Ukraine; however, the degree of volatility dropped later. The evidence is confirmed by the financial markets' low degree of volatility and persistence during this war situation.

In addition, the dynamic correlation results show that the correlation between the volatilities of different financial markets during the war was weaker than before, suggesting the occurrence of an impact. There is also evidence of the safe haven characteristic of the 10-Year Treasury Yield of the USA (TNX) of NY Mercantile. According to our findings, investors need to invest with special attention to the war between Russia and Ukraine and consider the US 10-year Treasury yield as one of their portfolio's assets.

## References

- Engle, R.: Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **20**(3), 339–350 (2002)
- Fischedick, M.: Energy supply risks, the energy price crisis and climate protection require joint responses. *Wirtschaftsdienst* **102**(4), 262–269 (2022)
- Gronwald, M.: Is bitcoin a commodity? On price jumps, demand shocks, and certainty of supply. *J. Int. Money Financ.* **97**, 86–92 (2019)
- Huther, M.: The problem of subjective value judgment: on the calculations of the cost of a Russian gas embargo. *Wirtschaftsdienst* **102**(4), 273–278 (2022)
- Maneejuk, P., Yamaka, W.: Significance test for linear regression: how to test without P-values? *J. Appl. Stat.* **48**(5), 827–845 (2021)
- Rotta, T.N., Parana, E.: Bitcoin as a digital commodity. *New Political Economy* (2022). <https://doi.org/10.1080/13563467.2022.2054966>



# Net Interest Margins of Vietnamese Commercial Banks: What Really Affects?

Nam Hai Pham<sup>(✉)</sup> and Thuy Kieu Thi Vo

Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam  
{namph, kieuvtt}@buh.edu.vn

**Abstract.** The study analyzes the factors affecting the net interest margins (NIM) of 30 commercial banks in Vietnam in the period 2012–2020 by using the Bayesian method via Gibbs sampling algorithm. The dependent variable in the research model is the NIM of Vietnamese commercial banks. The research model includes six independent variables: bank size, bank capital, operating costs, bank loans, inflation, and GDP growth. The research results show that the factors that positively affect the NIM of banks include bank size, bank capital, operating costs, outstanding loans, and inflation. Meanwhile, GDP growth has the opposite effect.

**Keywords:** Net interest margins · Commercial bank · Bayes

## 1 Introduction

Vietnam's economy has made remarkable progress in recent years with a high growth rate, and the economic structure has gradually shifted to industry and services. In particular, the role of the banking system is very important as the main channel for the capital of the economy in the context of the young and limited stock market. Similar to other emerging economies, the role of banks is indispensable and decisive for the development of the economy. In Vietnam, the main activities of commercial banks are mobilizing capital and lending, bringing the most significant profit for banks. However, at the same time, it is also a risky activity, causing unsafety for the banking system (Dang, 2021). And therefore, interest income from lending activities is an issue that banks always need to pay attention to in order to both ensure profits for the bank and ensure the ability to recover capital. NIM measures the difference between interest income and interest expenses that a bank can achieve through tight control of its earning assets and the pursuit of the lowest-cost sources of capital (Lestari et al., 2021). The higher this ratio, the better for the bank's business because, at this time, it earns more interest than it pays. If this ratio is low, the bank's profitable assets are not very profitable, or the bank has mobilized capital at high-interest rates, so the difference is low. NIM is an important measure of the bank's performance (Nguyen & Le, 2016). A bank's NIM also represents the health and safety of a bank's traditional operations of taking deposits and making loans (Angori et al., 2019). The fact that banks maintain low NIM may not fully cover the costs and cause damage to the bank. However, if banks maintain NIM too high, it can

lead to financial distress for borrowers, leading to insolvency and increasing the bank's risk of bankruptcy. Therefore, maintaining a reasonable NIM is an issue that banks need to consider cautiously but flexibly, depending on certain stages of the economy. In order to maintain a reasonable NIM, it is necessary to study the factors affecting NIM, thereby helping banks adjust NIM according to the characteristics of each specific bank.

In the world and in Vietnam, many studies have been carried out to evaluate the factors affecting the NIM of commercial banks. Typically, the studies of Ho and Saunders (1981), Allen (1988), Angbazo (1997), Demirgüç-Kunt and Huizinga (1999), or recent studies such as Angori (2019), Islam and Nishiyama (2016), Lestari et al. (2021), Nguyen and Pham (2018), Nguyen and To (2020), Dang (2021). These studies have partly elucidated the factors affecting the NIM of commercial banks in the world as well as in Vietnam. However, the studies were carried out with traditional methods. These methods have many limitations and controversies. Therefore, the new point of this study is that the author carried out the study according to the Bayesian method, which is a new, modern method and is highly appreciated by many researchers (Wang et al. 2019; Hung, 2020). The results of the study are the basis for commercial banks of the two groups to adjust NIM appropriately, contributing to improving operational efficiency and safety of Vietnamese commercial banks.

## 2 Literature Review

Ho and Saunders (1981) were the first to study the factors affecting the NIM of commercial banks, leading to many later studies on the NIM of commercial banks. In this study, the authors used data from 197 banks in the US during the period 1976–1979 and divided it into two steps. In the first step, independent variables representing bank characteristics are included in the model. They argue the existence of “pure arbitrage”, which is the price of providing instant services in the face of uncertainty caused by asynchronous deposit supply and loan demand. In the second stage. They attempted to measure pure arbitrage by looking at the number of imperfections and regulatory constraints. According to the research results of Ho and Saunders (1981), the NIM of commercial banks depends on the following factors: risk aversion, bank size, market structure, and fluctuations in credit and deposit interest rates. The limitation in the study of Ho and Saunders (1981) is that the authors do not consider the special role of commercial banks as providing intermediary financial services. To improve the model, Allen (1988) adds multiple types of loans with interdependent demand and concludes that NIM can be reduced when there is a demand elasticity between the bank's products. Next, Angbanzo (1997) added a default risk factor to the model.

Many empirical studies have been carried out by authors around the world in many different countries around the world. Typical is the study of Fungacova and Poghosyan (2011). In this study, the authors evaluated the impact of factors on the NIM of commercial banks in Russia from 1999 to 2007. Using the FEM method, the research results showed that the factors have a positive impact on NIM, including staffing costs and bank capital. The opposite factors include non-performing loans, bank size, liquid assets, Herfindahl index.

The study of Islam and Nishiyama (2016) aimed to understand the factors affecting NIM in South Asian countries (Bangladesh, India, Nepal, and Pakistan) period 1997 - to

2012. Applying FEM method, research results show that liquid assets, equity, required reserves, and operating costs are factors that have a positive impact on NIM of commercial banks. The factors of bank size, market power, and economic growth have negative effects on NIM.

Research by Lestari et al. (2021) on factors affecting NIM of commercial banks in Indonesia in 2015–2019. By GLS method, the authors conclude that the loan-to-deposit ratio and management efficiency has a positive effect on NIM. Meanwhile, bank size, credit risk, bank capital, and inflation have the opposite effect.

In Vietnam, Pham et al. (2019) studied the factors affecting the NIM of Vietnamese commercial banks. The study uses secondary data from 26 Vietnamese commercial banks in the period 2008 to 2017. Using the GLS method, the research results show that bank loans, capitalization, and inflation have a positive impact on NIM. Meanwhile, effective management has the opposite effect. In addition, bank size, credit risk, and loan-to-customer deposit ratio were not statistically significant.

Nguyen and Le (2016) study the factors affecting the NIM of 25 commercial banks in Vietnam from the period 2011 to 2014. By FEM method, the empirical results show that Bank capital, credit risk, and implicit interest cost are positively and statistically significant with NIM. Meanwhile, quality of management has a negative relationship with NIM.

## **Research Hypothesis**

### **Bank Size**

Large banks generally enjoy the trust of depositors and are therefore able to mobilize deposits at low-interest rates and lend at market rates. As a result, large banks have higher NIM than small banks. Therefore, the author hypothesizes:

**Hypothesis H1:** Bank size has a positive effect on NIM.

### **Bank Capital**

Banks with large capital have a good ability to withstand risks compared to other banks. Therefore, those banks can apply higher lending rates. In addition, banks need to maintain capital to meet safety standards in banking operations, and shareholders demand higher profits in proportion to the amount of capital contributed. On that basis, the author proposes the hypothesis:

**Hypothesis H2:** Bank capital has a positive effect on NIM.

### **Operating Costs**

The bank's operating expenses include staff salaries, advertising, etc. A bank's high operating costs mean that the bank is inefficient. To compensate for the high costs spent, banks need to impose high lending rates. Therefore, the author hypothesizes:

**Hypothesis H3:** Operating costs have a positive effect on NIM.

### **Bank Loans**

A bank with a high loan ratio will lead to high risk and income dependent on lending

activities. That bank will also have a lax credit management process, making it easy to accept loans. To offset potential risks in the future, that bank will impose a higher lending rate. Therefore, the author hypothesizes:

**Hypothesis H4:** Bank loans have a positive impact on NIM.

### **Inflation**

Inflation is often associated with market interest rates. Therefore, when inflation is high, banks will increase lending rates to cover expenses such as capital mobilization and operating costs, thereby increasing NIM. Therefore, the author hypothesizes:

**Hypothesis H5:** Inflation has a positive effect on NIM.

### **GDP Growth**

High GDP growth is often associated with expansionary monetary policy. During the period of high economic growth, the demand for deposits and loans increased. Due to the rapid increase in money supply, banks can reduce lending rates to increase competitiveness and increase the ability to bring capital into the economy. Due to the decrease in interest rates, the bank's NIM decreased accordingly. Therefore, the author proposes the hypothesis:

**Hypothesis H6:** GDP growth has a negative impact on NIM.

## **3 Data and Methodology**

### **3.1 Research Data**

The dataset used in the research is the unbalanced panel data of 30 Vietnamese commercial banks from 2012–2020. Out of the total of 30 banks, 17 are listed on the Ho Chi Minh City Stock Exchange, and Hanoi Stock Exchange and 13 are unlisted banks. The data is collected from the audited annual consolidated financial statements. Most banks today are developing in the direction of a group with many subsidiaries, so the separate financial statements do not fully reflect the financial position and business activities of the bank. Therefore, the new consolidated financial statements fully reflect the business performance of the bank.

### **3.2 Research Method and Models**

In this study, the author uses Bayesian regression method via the Gibbs sampling algorithm. Compared with previous studies on factors affecting the NIM of commercial banks, the Bayesian method is a new, modern approach and has many advantages over the traditional approach and is highly evaluated by many researchers (Thach et al., 2021; Wang et al., 2019; Hung, 2020; and Thach et al., 2019).

Previous studies on NIM of commercial banks were done by traditional methods, so information about the prior distribution of variables in the model is not available. Therefore, based on the information from the collected data set, the author proposes that the



normal distribution is the distribution of the regression coefficients of the variables and the Igamma distribution for the variances in the model. The specific a prior distribution is as follows:

$$\alpha \sim N(0; 100)$$

$$\sigma^2 \sim \text{Invgamma}(0.01; 0.01)$$

The research model is built based on previous studies by Fungacova and Poghosyan (2011), Islam and Nishiyama (2016) and some other authors, specifically as follows:

$$\text{NIM}_{it} = \beta_0 + \beta_1 \text{SIZE}_{it} + \beta_2 \text{CAP}_{it} + \beta_3 \text{OPE}_{it} + \beta_4 \text{LOAN}_{it} + \beta_5 \text{INF}_t + \beta_6 \text{GGDP}_t + u_{it}$$

where:

$\beta_0$ : constant

$\text{NIM}_{it}$ : NIM of bank  $i$  year  $t$ .

$\text{SIZE}_{it}$ : Size of bank  $i$  year  $t$ .

$\text{CAP}_{it}$ : capital of bank  $i$  year  $t$ .

$\text{OPE}_{it}$ : operating costs of bank  $i$  year  $t$ .

$\text{LOAN}_{it}$ : loans of bank  $i$  year  $t$ .

$u_{it}$ : error

## 4 Results and Discussion

### 4.1 Descriptive Statistics

The research sample is synthesized from financial statements and annual reports of 30 Vietnamese commercial banks from the period 2012 to 2020. The observations in the data set are made based on excluding inappropriate observations (Table 1). The descriptive statistics of the variables in the model are presented in Table 2.

The results from Table 2 show that the average NIM of banks is 2.92%, the highest and lowest values are 10.49% and 0.43% respectively. The average value of bank size, bank capital, operating expenses, and bank loans is 32.43; 9.37%; 1.71%; 56.85% respectively.

**Table 1.** Definition of the variables

Var	Notation	Previous studies	Expected results	Formula
Dependent	NIM	Fungacova and Poghosyan (2011), Islam and Nishiyama (2016), Pham et al. (2019), Lestari et al. (2021),		$NIM = \frac{\text{Interest income} - \text{Interest expense}}{\text{Total earning assets}}$
Independent	SIZE	Fungacova and Poghosyan (2011), Islam and Nishiyama (2016), Lestari et al. (2021), Nguyen and To (2020)	+	The logarithm of total assets
	CAP	Nguyen and To (2020), Fungacova and Poghosyan (2011), Islam and Nishiyama (2016),	+	Bank capital/Total assets
	OPE	Islam and Nishiyama (2016)	+	Operating costs/Total assets
	LOAN	Nguyen and Le (2016), Maudos and De Guevara (2004), Zhou and Wong (2008)	+	Total loans/Total assets
	INF	Lestari et al. (2021), Nguyen and Pham (2018)	+	Yearly inflation rate
	GGDP	Islam and Nishiyama (2016)	-	Yearly GDP growth rate

## 4.2 Regression Results and Discussion

Convergence diagnosis of Markov Chain Monte Carlo (MCMC) chains is performed to ensure the model is robust and Bayesian inference is reasonable (Nguyen, 2020). Convergence of MCMC chains is diagnosed through trace plot, posterior distribution plot, and autocorrelation plot.

**Table 2.** Descriptive statistics

Var.	Obs	Mean	Std.dev	Min	Max
NIM	247	0.0292	0.0147	0,0043	0,1049
SIZE	248	32.4213	1.1431	29.4391	34,9553
CAP	248	0.0937	0.0621	0.0262	0.6140
OPE	248	0.0171	0.0068	0.0067	0.0692
LOAN	248	0.5685	0.1134	0.2162	0.7880
INFLAT	261	0.0422	0.0230	0,0063	0,0921
GGDP	261	0.0592	0.0123	0.0291	0.0708

Source Results from Stata software.

**Table 3.** Summary of regression results for the dependent variable NIM

	Mean coefficient	Std.	MCSE	95% Cred. Interval	
SIZE	0.0021	0.0008	0.0000	0.0005	0.0038
CAP	0.0218	0.0210	0.0002	-0.0194	0.0627
OPE	1.7462	0.1689	0.0016	1.4116	2.0769
LOAN	0.0127	0.0072	0.0000	-0.0012	0.0271
INF	0.0039	0.0361	0.0003	-0.0374	0.1057
GGDP	-0.0334	0.0637	0.0006	-0.1559	0.0925
con_	-0.0789	0.0280	0.0002	-0.1347	-0.0245
var	0.0001	0.0000	0.0000	0.0001	0.0001

Source Results from Stata software.

The test results from Fig. 1 show that the histograms do not create trends, the autocorrelation histogram shows low autocorrelation, and the shape of the histograms simulating the shape of the probability distributions is uniform. Therefore, it can be concluded that Bayesian inference is reasonable, and the results can be used for analysis.

Table 3 shows that the variables that positively affect the NIM of Vietnamese commercial banks are bank size, bank capital, operating costs, bank loans, and inflation. Meanwhile, GDP growth has a negative impact on NIM.

Bank size (SIZE) has a positive effect on NIM, consistent with the research hypothesis and research results of Nguyen and To (2021) but contrary to the results of Fungacova and Poghosyan (2011), Islam and Nishiyama (2016), Lestari et al. (2021). Large-scale banks receive the trust of depositors, so the cost of capital mobilization is lower than that of small-sized banks. Larger banks can lend at higher interest rates due to their large customer base and the ability to effectively reach customers.

Bank capital (CAP) has a positive effect on the NIM of Vietnamese commercial banks. Banks need to attract investors to increase capital and meet regulations on safety

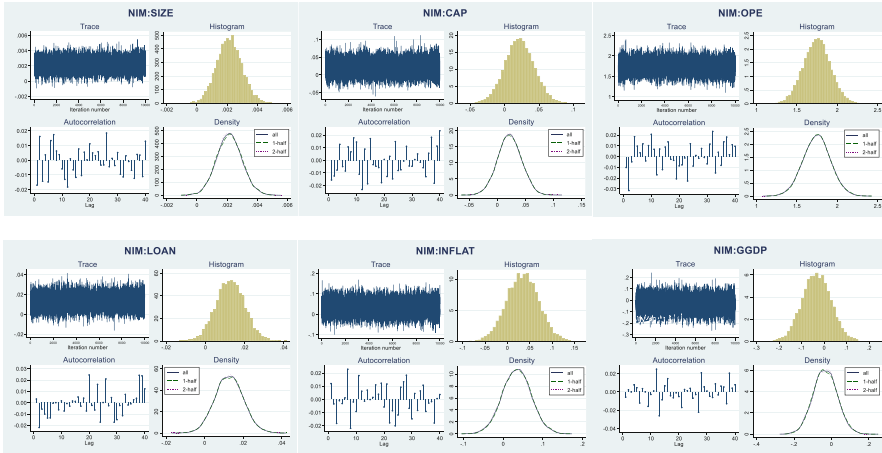


Fig. 1. Graphical diagnostics for MCMC convergence

in banking operations. Therefore, banks need to increase lending rates to achieve higher income, leading to an increase in their NIM. This result is consistent with the research hypothesis and studies of Fungacova and Poghosyan (2011), Islam and Nishiyama (2016), Nguyen and Do (2014), Nguyen and Pham (2016), Dang (2021), Nguyen and To (2020), but contrary to the results of Lestari et al. (2021).

Operating costs (OPE) have a positive impact on the bank’s NIM, consistent with the research hypothesis and research of Islam and Nishiyama (2016). When banks control operating costs less effectively, they need to compensate with higher lending rates. In other words, borrowers need to pay higher interest rates to banks that are less efficient in terms of costs.

Bank loans (LOAN) have a positive impact on the NIM of banks, similar to the research results of Nguyen & Le (2016), Maudos and De Guevara (2004), Zhou and Wong (2008) and research hypothesis. Banks with a high loan to total assets rate tend to accept higher risks and therefore need to increase lending rates to offset the risk.

Inflation has a positive impact on the NIM of Vietnamese commercial banks. This result is consistent with the research hypothesis and the studies of Lestari et al. (2021), Nguyen and Pham (2018). When inflation rises, banks increase lending rates to limit the impact of inflation on bank earnings, leading to an increase in NIM.

GDP growth has a negative impact on the NIM of Vietnamese commercial banks. When economic growth is high, banks attract more deposits from customers. Therefore, in order to effectively use the mobilized money, banks need to reduce lending rates to increase competitiveness, leading to a decrease in NIM. This result is consistent with the research hypothesis and research of Islam and Nishiyama (2016).

## 5 Conclusions and Policy Implications

The study was conducted to evaluate the factors affecting the NIM of listed and unlisted banks on the stock market in Vietnam. Using the data set of 17 listed banks and 13 unlisted banks in the period 2012–2020, by Bayesian method via Gibbs sampling algorithm, the results show that the factors that have a positive impact on the NIM of Vietnamese commercial banks are bank size, bank capital, operating costs, bank loans, and inflation. The factor that has a negative impact on NIM is GDP growth. From the above results, the author proposes some solutions for banks as follows:

Firstly, banks need to improve the efficiency of controlling operating costs. When banks increased lending rates to compensate for the weakness in cost management, it negatively impacted borrowers, increasing the burden of paying interest. As a result, non-performing loans will increase, and the loan lost provision will also increase accordingly and reduce the bank's profit.

Second, banks need to diversify their business activities and revenue sources and avoid relying too much on lending activities. From there, it will be possible to reduce lending interest rates for borrowers, increasing the bank's operational efficiency.

Third, banks can raise capital through various sources such as retained earnings or the issue of convertible bonds. The increase in capital from issuing shares to shareholders puts the bank under pressure to increase profits, mainly by increasing lending interest rates and causing an interest payment burden on customers.

## References

- Allen, L.: The determinants of bank interest margins: a note. *J. Financ. Quant. Anal.* **23**(2), 231–235 (1988)
- Angbazo, L.: Commercial bank net interest margins, default risk, interest-rate risk and off-balance sheet banking. *J. Bank. Finance* **21**, 55–87 (1997)
- Angori, G., Aristei, D., Gallo, M.: Determinants of banks' net interest margin: Evidence from the Euro area during the crisis and post-crisis period. *Sustainability*, **11**, 2–20 (2019)
- Demirguc-Kunt, A., Huizinga, H.: Determinants of commercial bank interest margins and profitability: Some international evidence. *World Bank Econ. Rev.* **13**, 379–408 (1999)
- Dang, T.L.P.: Factors affecting the net interest margins of Vietnamese commercial banks. *J. Finance*, March 2021 (2021)
- Fungáčová, Z., Poghosyan, T.: Determinants of bank interest margins in Russia: Does bank ownership matter? *Econ. Syst.* **35**(4), 481–495 (2011)
- Hung, N.T.: On the calculus of subjective probability in behavioral economics. *Asian J. Econ. Bank.* **4**(1) (2020)
- Ho, T.S., Saunders, A.: The determinants of bank interest margins: Theory and empirical evidence. *J. Financ. Quant. Anal.* **16**(4), 581–600 (1981)
- Islam, M.S., Nishiyama, S.-I.: The determinants of bank net interest margins: A panel evidence from South Asian countries. *Res. Int. Bus. Financ.* **37**, 501–514 (2016). <https://doi.org/10.1016/j.ribaf.2016.01.024>
- Lestari, H.S., Chintia, H., Akbar, I.C.: Determinants of net interest margin on conventional banking: Evidence in Indonesia stock exchange. *Jurnal Keuangan dan Perbankan* **25**(1), 104–116 (2021)
- Maudos, J., De Guevara, J.F.: Factors explaining the interest margin in the banking sectors of the European Union. *J. Bank. Finance* **28**(9), 2259–2281 (2004)

- Nguyen, N.T.: How to explain when the ES is lower than one? A Bayesian nonlinear mixed-effects approach. *J. Risk Financ. Manag.* **13**(2), 1–17 (2020). <https://doi.org/10.3390/jrfm13020021>
- Nguyen, D.A., To, T.H.G.: Factors affecting the net interest margins of Vietnamese joint stock commercial banks. *J. Finance*, December 2020 (2021)
- Nguyen, A.T., Pham, T.N.: Net interest margins of Vietnamese commercial banks in the period 2005–2017 - An empirical study. *J. Bank.* July 2019 issue (2018)
- Nguyen, T.B.T., Le, M.T.: Research on factors affecting the net interest margins of commercial banks in Vietnam. *J. Sci.* **20**, 43–48 (2016)
- Nguyen, K.T., Do, T.T.H.: Analysis of factors affecting the net interest margins of Vietnamese commercial banks. *Sci. J. Vietnam Natl. Univ. Hanoi Econ. Bus.* **4**, 55–65 (2014)
- Pham, A.H., Tran, C.K.Q., Vo, L.K.T.: Determinants of net interest margins in Vietnam banking industry. In: Kreinovich, V., Thach, N.N., Trung, N.D., Van Thanh, D. (eds.) *ECONVN 2019*. SCI, vol. 809, pp. 417–426. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_30](https://doi.org/10.1007/978-3-030-04200-4_30)
- Thach, N.N., Anh, L.H., An, P.T.H.: The effects of public expenditure on economic growth in Asia countries: A bayesian model averaging approach. *Asian J. Econ. Bank.* **3**(1), 126–149 (2019)
- Ngoc, N.T., Kreinovich, V., Trung, N.D., (eds.): *Data Science for Financial Econometrics*. Studies in Computational Intelligence SCI, vol. 898 (2021). Springer Cham. <https://doi.org/10.1007/978-3-030-48853-6>
- Wang, C., Wang, T., Trafimow, D., Chen, J.: Extending a priori procedure to two independent samples under skew normal settings. *Asian J. Econ. Bank.* **3**(2), 29–40 (2019)
- Zhou, K., Wong, M.C.: The determinants of net interest margins of commercial banks in mainland China. *Emerg. Mark. Financ. Trade* **5**(44), 41–53 (2008)



# Credit Growth: An Investigation of Vietnamese Commercial Banks

Nam Hai Pham<sup>1</sup>(✉), Nguyen Ngoc Thach<sup>1</sup>, Tuan Van Ngo<sup>1</sup>, and Tri Minh Hoang<sup>2</sup>

<sup>1</sup> Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam

{namph, thachnn, tuannv}@buh.edu.vn

<sup>2</sup> HUTECH University, Ho Chi Minh City, Vietnam

hm.tri@hutech.edu.vn

**Abstract.** This study assesses the factors affecting the credit growth of Vietnamese commercial banks in the period 2012–2020. The study uses the Bayesian regression model via the Metropolis-Hastings sampling algorithm to estimate the results based on annual panel data of 17 commercial banks listed on the Vietnamese stock market. The research results show that the factors that positively impact credit growth include deposit growth, net interest margins ratio (NIM), and GDP growth. Factors that have the opposite effect are non-performing loans, bank size, and inflation. From the findings of this study, bank administrators will have a basis to decide the appropriate credit growth rate and add value to the bank's shareholders.

**Keywords:** commercial banks · credit growth · listed banks

## 1 Introduction

Bank credit in a country can affect economic growth and the effectiveness of monetary policy in that country (Jessica & Chalid, 2019). Therefore, credit growth is an issue that policymakers are always interested in and have closely directed because it is an essential driving force that represents the investment and consumption needs of the economy. In addition, the role of credit is critical because it is a source of funding for production expansion, supplementing enterprises' working capital and the population's consumption needs. Especially in Vietnam, where the stock market is still young and limited, unable to meet the economy's medium and long-term capital needs, the role of credit becomes even more critical (Batten & Vo, 2019). Banks also enjoy many benefits from credit growth, such as increased profits, expanded market share, and the ability to expand the provision of banking services and products to customers.

On the other hand, rapid credit growth can have a negative effect on the economy because banks can be lenient in credit granting, leading to deterioration in credit quality, increasing bankruptcy risk, and increasing credit risk (Awdeh, 2017; Igan & Pinheiro, 2011). Researchers have also found evidence showing a positive relationship between credit growth and bank risk (Bhowmik & Sarker, 2021; Laeven & Majnoni, 2003). To

increase profits to satisfy the desires of shareholders, banks may lower the standards to grant more credit, increasing credit risk and other risks in the future. Therefore, policymakers can use credit growth data to predict the economy's health. Excessive credit growth can lead to a financial crisis, while a slowdown in credit growth can signal that the economy is in recession (Awdeh, 2017).

Therefore, credit growth that is too high or too low has bad effects on the economy in general and commercial banks in particular. Moderate credit growth can promote healthy economic growth and the safety of banks, becoming a driving force for other areas to develop. Therefore, maintaining a reasonable credit growth rate by controlling factors affecting credit growth can help banks achieve appropriate profits and operate safely while also helping the economy grow. Many studies have been conducted in the world and Vietnam on the factors affecting the credit growth of commercial banks. These studies were carried out by FEM, REM, and GMM methods. These methods can produce controversial results. Therefore, the novelty of this study is that the study was carried out using the Bayesian method. This new and modern method is highly appreciated by researchers, giving reliable results (Khrennikova, 2019; Galindo et al. 2020; Thach et al. 2019). The results of the study are the basis for policymakers to adjust credit growth and increase the operational efficiency and safety of Vietnam's banking system.

## 2 Literature Review

Credit plays an essential role in the economy's circulation of money and goods. Besides this vital role, excessive credit growth also causes an imbalance between money and goods, resulting in high inflation. Therefore, one of the indicators that monetary policy operators are interested in is the credit growth rate of the commercial banking system, which reflects the credit of the commercial banking system in the present compared to the past (Ivanović, 2016). Empirical studies on the factors affecting credit growth have been carried out by many researchers in the world and Vietnam, contributing to the understanding of this topic in many regions and countries (Ivanović, 2016; Awdeh, 2017; Pasaribu and Mindosa, 2021; Phan and Tran, 2021; Kowalska et al., 2019). In Asia, Pasaribu and Mindosa (2021) explore the factors affecting the credit growth of commercial banks in Indonesia from 2002 to 2018. Using GMM method and a dataset of 86 commercial banks in Indonesia, the research results show that credit growth last year, customer deposit growth, and non-performing loan ratio are the factors affecting the credit growth of commercial banks in Indonesia.

Al-Shammari and El-Sakka (2018) study macro factors affecting credit growth of OECD countries in the period 2011–2013. Research results show that exchange rate, national debt, money supply, interest rates, inflation, GDP growth, and fixed capital formation affect the credit growth of OECD countries.

In Europe, Kowalska et al. (2019) studied the impact of bank capital on the credit growth of commercial banks in Poland from 2000 to 2012. Using quarterly data set of 237 banks and methods FEM regression, the results show that bank capital positively impacts credit growth. Besides, the control variables are NIM, customer deposits, interbank interest rates, and unemployment rate, which all impact the credit growth of banks in Poland.



Ivanović (2016) studies the factors affecting the credit growth of commercial banks in Montenegro. Using the data set of 11 commercial banks in the period 2004–2014 and the FEM method, the results show that the non-performing loan ratio, customer deposit growth, and GDP growth are the factors affecting the credit growth of commercial banks in Montenegro.

In the Middle East, Awdeh (2017) explores the factors affecting the credit growth of commercial banks in Lebanon from 2000 to 2015. Using FEM method and data from 34 commercial banks, research result shows that customer deposit growth, GDP growth, inflation, and money supply are factors that have a positive impact on credit growth. On the contrary, the non-performing loan ratio, lending interest rates, government bond interest rates, public debt, and remittance are the factors that negatively affect the credit growth of commercial banks in Lebanon.

In Vietnam, research by Phan and Tran (2021) shows that customer deposit growth, liquidity, and GDP growth are factors that have a positive impact on credit growth. The factors of non-performing loan ratio, ROE, inflation, and interest rates are the factors that have opposite effects.

Pham (2017) explores factors affecting the credit growth of Vietnamese commercial banks in the period 2014 to 2017. Research results by REM and GMM methods show that customer deposit growth and GDP growth are two factors that have a positive impact on helping banks to grow credit. In contrast, the non-performing loan ratio, bank size, and inflation rate are factors that have a negative impact on the credit growth of banks.

Research results by Nguyen (2021) on factors affecting credit growth of 16 commercial banks listed on Vietnam's stock market in the period 2011–2020 show that credit size in the previous period and growth of deposits have a positive influence on credit growth of Vietnamese commercial banks. In contrast, bank size and non-performing loan ratio have a negative effect.

## Research hypothesis

### Customer Deposit Growth

Commercial banks depend on customer deposits to provide credit (Cornett et al., 2011). When customer deposit growth is high, banks can increase their capital for lending, leading to higher credit growth (Ivanovic, 2016). In addition, when customer deposits increase, banks have a greater incentive to find customers to borrow and achieve higher profits. Therefore, the author hypothesizes:

**Hypothesis H1:** customer deposit growth has a positive impact on credit growth.

### Non-performing Loan Ratio

Banks with high NPL ratios were forced to readjust their operations to control credit quality, leading to a decline in credit growth. Once the Non-performing loan ratio is high, banks have difficulty mobilizing capital and developing products and services. As a result, banks are forced to narrow their credit activities (Pham, 2017). Therefore, the author hypothesizes:

**Hypothesis H2:** NPL ratio has a negative impact on credit growth.

### **Bank Size**

Large-scale banks have many customers, so there are many advantages for banks to develop loans and mobilize capital. However, due to a large number of customers, large-scale banks have many options and always consider carefully before deciding to grant credit for fear of facing many risks of not being able to recover capital when payment is due. In contrast, small-sized banks have a smaller number of customers, so for profit, they often promote lending without careful consideration (Nguyen, 2021; Pham, 2017). Therefore, the credit growth rate of large banks may be lower than that of small banks. On that basis, the author hypothesizes:

**Hypothesis H3:** bank size has a negative effect on credit growth.

### **Net Interest Margins (NIM)**

NIM is an indicator of the bank's profit from lending. The NIM is the percentage difference between a bank's interest income and interest expense. The NIM shows how much banks enjoy the interest rate differential between capital mobilization and credit activities (Kowalska et al., 2019). When a bank achieves a high NIM, it will have more incentive to extend credit and increase credit growth. Therefore, the author hypothesizes:

**Hypothesis H4:** NIM has a positive impact on credit growth.

### **Inflation**

Inflation can affect a country's monetary policy (Al-Shammari & El-Sakka, 2018). Governments tend to tighten monetary policy to control inflation when inflation is high. That makes banks forced to be more restrictive in credit granting and slows down credit growth. Therefore, the author hypothesizes:

**Hypothesis H5:** inflation has a negative impact on credit growth.

### **GDP Growth**

High GDP growth will create more development opportunities for businesses. Besides, people's income will be higher, and consumption will increase. Therefore, GDP Growth will boost investment and consumption demand. So loan demand will also increase to meet these needs, leading to higher credit growth. Therefore, the author proposes the hypothesis:

**Hypothesis H6:** GDP growth has a positive impact on credit growth.

## **3 Methodology and Model**

### **3.1 Research Data**

Research using data from 17 commercial banks listed on the Vietnam stock market. Data sources are collected at the end of each year in the period 2012–2020 from 17 banks. The indicators used in the study to measure the model's variables are secondary data published in the audited financial statements and annual reports of commercial banks. Macroeconomic indicators each year are collected from the General Statistics Office of Vietnam during the same period.

### 3.2 Research Method and Model

Previous studies on factors affecting banks' credit growth have been conducted using the frequency method. However, this method's testing of statistical hypotheses reveals many limitations in interpreting results and forecasting in many cases (Thach et al., 2022; Briggs & Hung, 2019; Hung, 2019). Meanwhile, the Bayesian method has many outstanding advantages over the frequency method and is highly appreciated by researchers. Besides, with the development of data science and computers' computing power, complex Bayesian statistics algorithms have been handled quickly and easily. Therefore, the Bayesian approach is more and more widely used, especially in social sciences and medicine. Therefore, this study will be carried out according to the Bayesian approach. Bayesian statistics are based on the assumption that the model parameters are random numbers. Then, with the Bayesian probabilistic rule, this analysis provides a way to combine prior information with evidence from collected data to generate posterior distributions of model parameters. Because the studies on the factors affecting banks' credit growth are conducted by the frequency method, information on the prior distribution of the variables in the research model is unavailable. Therefore, in this study, the author uses the normal distribution of  $N(0,100)$  for the observed variables and the Igamma distribution  $(0.01; 0.01)$  for the variances in the model (Table 1, 2 and 3).

Based on the research model of Pasaribu and Mindosa (2021), Ivanović (2016), and some other studies, the author proposes a research model as follows:

$$GCREDIT_{i,t} = \alpha_0 + \alpha_1 GDEP_{i,t} + \alpha_2 NPL_{i,t} + \alpha_3 SIZE_{i,t} + \alpha_4 NIM_{i,t} + \alpha_5 INFLAT_t + \alpha_6 GGDP_t + \varepsilon_{i,t}$$

where:

$GCREDIT_{i,t}$ : credit growth rate of bank  $i$  year  $t$

$\alpha_0$ : constant

$GDEP_{i,t}$ : customer deposit growth rate of bank  $i$  year  $t$

$NPL_{i,t}$ : non-performing loan ratio of bank  $i$  year  $t$

$SIZE_{i,t}$ : bank size of bank  $i$  year  $t$

$NIM_{i,t}$ : net interest margins ratio of bank  $i$  year  $t$

$INFLAT_t$ : yearly inflation rate

$GGDP_t$ : yearly GDP growth rate

$\varepsilon_{i,t}$ : error

**Table 1.** Variable definitions

Variables		Notation	Previous studies	Expected results	Proxy variables
<i>Dependent variable</i>	credit growth	GCREDIT	Ivanović (2016), Awdeh (2017), Pasaribu and Mindosa (2021), Phan and Tran (2021), Kowalska et al. (2019)		$(\text{Credit}_t - \text{Credit}_{t-1})/\text{Credit}_{t-1}$
<i>Independent variables</i>		GDEP	Ivanović (2016), Awdeh (2017), Pasaribu and Mindosa (2021), Phan and Tran (2021), Pham (2017), Nguyen (2021)	+	$(\text{Deposit}_t - \text{Deposit}_{t-1})/\text{Deposit}_{t-1}$
		NPL	Nguyen (2021), Ivanović (2016), Phan and Tran (2021), Awdeh (2017)	–	Non-performing loans to total loans ratio
		SIZE	Nguyen (2021), Pham (2017)	–	Logarithm of bank total assets
		NIM	Kowalska et al. (2019)	+	$\frac{\text{Interestincome} - \text{Interestexpense}}{\text{Total earning assets}}$

*(continued)*

**Table 1.** (continued)

Variables	Notation	Previous studies	Expected results	Proxy variables
	INFLAT	Awdeh (2017), Pham (2017), Phan and Tran (2021), Al-Shammari và El-Sakka (2018)	–	Yearly inflation rate
	GGDP	Awdeh (2017), Ivanović (2016), Phan and Tran (2021), Pham (2017),	+	Yearly GDP growth rate

**Table 2.** Descriptive statistics of research variables in the period 2012–2020

Var.	Number of obs.	Mean	Std.	Min	Max
GCREDIT	136	0.2116	0.1628	–0.1390	1.0618
GDEP	136	0.1804	0.1471	–0.0803	0.8269
NPL	140	0.0217	0.0128	0.0047	0.0900
SIZE	153	32.9502	0.9216	30.3470	34.9553
NIM	153	0.0305	0.0127	0.0043	0.0884
INFLAT	153	0.0422	0.2309	0.0063	0.0921
GGDP	153	0.0592	0.0124	0.0291	0.0708

Source: Calculation results from STATA software.

## 4 Research Results and Discussion

### 4.1 Descriptive and Correlation Statistics

### 4.2 Results

Table 4 presents the results of multivariate regression with panel data by Bayesian method via Random-walk Metropolis-Hastings algorithm and Gibbs sampling.

To ensure that the model is robust and that Bayesian inference based on MCMC chains simulation (Markov Chain Monte Carlo) is reasonable, the author performs a convergence diagnosis of MCMC chains.

**Table 3.** Correlation matrix

Var.	GCREDIT	GDEP	NPL	SIZE	NIM	INFLAT	GGDP
GCREDIT	1.0000						
GDEP	0.5916	1.0000					
NPL	-0.0144	0.1227	1.0000				
SIZE	-0.1466	-0.1192	-0.3301	1.0000			
NIM	0.0171	-0.1249	-0.0037	0.1333	1.0000		
INFLAT	-0.0422	0.0894	0.2698	-0.0975	0.0658	1.0000	
GGDP	0.1041	-0.0447	0.0199	0.0871	-0.0554	-0.2048	1.0000

Source: Calculation results from STATA software.

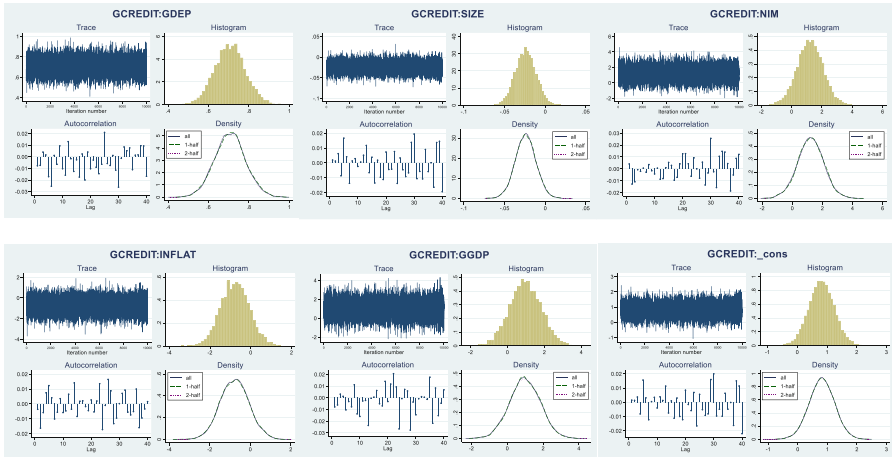
**Table 4.** Summary of regression results for the dependent variable CCREDIT

	Mean coefficient	Std.	MCSE	95% Cred. Interval	
GDEP	0.6225	0.0744	0.0007	0.4722	0.7717
NPL	-1.4689	0.9643	0.0094	-3.3596	.04039
SIZE	-0.0194	0.0124	0.0001	-0.0438	0.0049
NIM	1.2059	0.7952	0.0076	-0.3391	2.7842
INFLAT	-0.4045	0.7008	0.0078	-1.7562	1.0040
GGDP	1.3043	0.7990	0.0079	-0.2417	2.8790
_cons	0.6656	0.4279	0.0042	-0.1719	1.5107
var	0.0129	0.0017	0.0000	0.0101	0.0168

Source: Authors' calculation.

This diagnosis can be made through trace plot, autocorrelation plot, and histogram. The results of the convergent diagnostics of the MCMC chains are shown in Fig. 1. The results of the convergence diagnostics of the MCMC chains are shown in Fig. 1. The trace plots and autocorrelation plots from Fig. 1 show the autocorrelation in the low. Histogram and density plots both show the simulation of the normal distribution shape of the parameters in the model. From the results of testing the convergence of MCMC chains, it can be concluded that the MCMC chains have converged to a stationary distribution. Therefore, it can be concluded that Bayesian inference is robust, and the results can be used for analysis (Nguyen, 2020),

The results from Table 4 show that customer deposit growth (GDEP) has a positive impact on the credit growth of commercial banks, consistent with the research hypothesis and studies of Ivanović (2016), Awdeh (2017), Pasaribu and Mindosa (2021), Phan and Tran (2021), Pham (2017), Nguyen (2021). Banks with a good ability to mobilize customer deposits will have higher ability and motivation to find customers to extend credit, leading to higher credit growth.



**Fig. 1.** Graphical diagnostics for MCMC convergence

NPL ratio (NPL) has a negative impact on credit growth (GCREDIT) of Vietnamese commercial banks, consistent with the research hypothesis and studies of Nguyen (2021), Ivanović (2016), Phan and Tran (2021), Awdeh (2017). Banks with high NPL ratios tend to reduce credit growth to better control risk. In addition, a high NPL ratio will hinder the credit activities of commercial banks due to tighter control by the State Bank of banks with poor credit quality.

Consistent with the research hypothesis, bank size (SIZE) negatively impacts commercial banks' credit growth. This means that the larger the bank, the lower the credit growth. Small-sized banks tend to promote higher credit activities in search of profits. Unlike large-scale banks, small-scale banks are in the development stage, so their activities depend mainly on credit, so the credit growth rate is also higher than that of large banks. This result is also similar to the study results of Nguyen (2021) and Pham (2017).

Net interest margins (NIM) positively impacts commercial banks' credit growth, consistent with the research hypothesis. Banks with higher NIM will earn more profit and promote stronger credit activity. This result is also similar to the study results of Kowalska et al. (2019).

Inflation (INFLAT) has a negative impact on credit growth. When inflation increases, banks will have more difficulty mobilizing deposits and granting credit, negatively affecting the credit growth of commercial banks. Moreover, the government will tighten monetary policy to curb inflation, reducing the credit demand of the economy. This result is consistent with the research hypothesis and studies of Pham (2017), Phan and Tran (2021), Al-Shammari, and El-Sakka (2018) but contrary to the research results of Awdeh (2017).

GDP growth (GGDP) has a positive impact on the credit growth of Vietnamese commercial banks, consistent with the research hypothesis and studies of Awdeh (2017), Ivanović (2016), Phan and Tran (2021), Pham (2017). When the economy has a high growth rate, businesses operate better and have more credit needs to expand production and business activities, leading to increased credit growth.

## 5 Conclusion and Policy Implications

The study was carried out to assess the factors affecting the credit growth of Vietnamese commercial banks in the period 2012–2020. By the Bayesian method via Metropolis-Hastings algorithm, the research results show that factors that positively impact credit growth include deposit growth, NIM, and GDP growth. The opposite factors are NPL ratio, bank size, and inflation. On that basis, the study proposes a number of solutions for commercial banks to adjust credit growth appropriately, specifically as follows:

Firstly, deposit growth has a positive impact on the credit growth of Vietnamese commercial banks. However, in the structure of customer deposits, short-term deposits account for a large proportion. Meanwhile, the economy's medium and long-term credit demand is very high. That could lead banks to face liquidity risk. Therefore, commercial banks need to increase the proportion of medium and long-term deposits in total deposits and the proportion of short-term credit to reduce risks in the future.

Second, NPLs is a factor that has a negative impact on the credit growth of commercial banks. Therefore, commercial banks need to better control NPLs by perfecting the credit process, stricter credit appraisal, and not lowering credit standards. In particular, commercial banks need to limit credit granting to potentially risky sectors such as real estate and securities and focus on lending to production and business sectors. In addition, it is necessary to actively focus on recovering and handling NPLs, aiming to gradually reduce the NPL ratio to below the safe level of 1%.

Thirdly, banks with small size and high credit growth may face huge credit risks due to poor credit quality and credit portfolios focusing too much on risky areas. Small-sized banks need to focus on credit quality instead of credit growth.

Fourth, banks with high NIM can accelerate credit growth. However, customers may experience the burden of paying interest if the bank's NIM is too high. Therefore, banks with high NIM can reduce interest rates for customers, thereby improving debt repayment capacity, reducing NPL ratio, and credit growth will be more sustainable in the future.

## References

- Al-Shammari, N., El-Sakka, M.: Macroeconomic determinants of credit growth in OECD countries. *Int. J. Bus.* **23**(3), 217–234 (2018)
- Awdeh, A.: The determinants of credit growth in Lebanon. *Int. Bus. Res.* **10**(2), 9–19 (2017)
- Batten, J., Vo, X.V.: Determinants of bank profitability – Evidence from Vietnam. *Emerg. Mark. Financ. Trade* **55**(1), 1–12 (2019)
- Bhowmik, P.K., Sarker, N.: Loan growth and bank risk: empirical evidence from SAARC countries. *Heliyon* **7**(5), 1–10 (2021)
- Briggs, W., Nguyen, H.T.: Clarifying ASA's view on P-values in hypothesis testing. *Asian J. Econ. Bank.* **3**(2), 1–16 (2019)
- Cornett, M.M., McNutt, J.J., Strahan, P.E., Tehranian, H.: The liquidity risk management and credit supply in the financial crisis. *J. Financ. Econ.* **101**(2), 297–312 (2011)
- Galindo, O., Svitek, M., Kreinovich, V.: Quantum (and more general) models of research collaboration. *Asian J. Econ. Bank.* **4**(1) (2020)
- Hung, T.N.: Toward improved models for decision making in economics. *Asian J. Econ. Bank.* **3**(1), 1–19 (2019)



- Igan, D., Pinheiro, M.: Credit Growth and Bank Soundness: Fast and Furious? IMF Working Paper WP/11/278, International Monetary Fund, Washington D.C (2011)
- Ivanović, M.: Determinants of credit growth: the case of montenegro. *J. Cent. Bank. Theory Pract.* **5**(2), 101–118 (2016)
- Jessica, T., Chalid, A.: Determinants of bank loans in Indonesia. *Adv. Soc. Sci. Educ. Humanit. Res.* **558**, 505–512 (2019)
- Khrennikova, P.: Quantum probability based decision making in finance: from individual preferences to market outcomes. *Asian J. Econ. Bank.* **3**(1) (2019)
- Kowalska, I., Olszak, M.: Bank Competition and the Effects of Macroprudential Policy on Pro-cyclicality of Lending. UW Faculty of Management Working Paper Series, No 8/2019 (2019). Available at SSRN: <https://ssrn.com/abstract=3471170>. <https://doi.org/10.2139/ssrn.3471170>
- Olszak, M., Kowalska, I., Świtąła, F.: Determinants of loans growth in cooperative banks in Poland: Does capital ratio matter? In: Jajuga, K., Locarek-Junge, H., Orłowski, L.T., Staehr, K. (eds.) *Contemporary Trends and Challenges in Finance*. SPBE, pp. 25–37. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-15581-0\\_3](https://doi.org/10.1007/978-3-030-15581-0_3)
- Laeven, L., Majnoni, G.: Loan loss provisioning and economic slowdowns: too much, too late? *J. Financ. Intermediation* **12**, 178–197 (2003)
- Nguyen, V.T.: Factors affecting credit growth of Vietnamese commercial banks. *J. Financ.* (2021). November 2021 issue
- Pasaribu, P., Mindosa, B.: The bank specific determinants of loan growth and stability: Evidence from Indonesia. *J. Indonesian Econ. Bus.* **36**(2), 93–123 (2021)
- Pham, X.Q.: Factors affecting credit growth of Vietnamese commercial banks in the period 2007–2014. *J. Sci.* (28), 40–47 (2017)
- Phan, T.H.Y., Tran, H.Y.: Factors affecting credit growth of Vietnamese commercial banks in the period 2014–2019. *J. Monetary Fin. Markets* (2021). Issue 13/2021
- Thach, N.N., Anh, L.H., An, P.T.H.: The effects of public expenditure on economic growth in Asia countries: A bayesian model averaging approach. *Asian J. Econ. Bank.* **3**(1), 126–149 (2019)
- Thach, N.N., Kreinovich, V., Ha, D.T., Trung, N.D.: *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty, and Related Topics*. Springer, Cham, vol. 427 (2022). <https://doi.org/10.1007/978-3-030-98689-6>



# Forecasting the Exchange Rate for the Thai Baht Against the Chinese Yuan by Using a Genetic Algorithm-Based Subset Autoregressive Integrated Moving Average Model

Tassathorn Poonsin, Vayu Thanomsing, Thanakorn Thunjang,  
and Worrawate Leela-apiradee<sup>(✉)</sup>

Department of Mathematics and Statistics, Faculty of Science and Technology,  
Thammasat University, Pathum Thani 12121, Thailand  
[worrawate@mathstat.sci.tu.ac.th](mailto:worrawate@mathstat.sci.tu.ac.th)

**Abstract.** Accurate forecasting of foreign exchange rates plays a crucial role in future global financial market investment, international business decision-making, and travel planning. This paper proposes a model for forecasting the daily exchange rate for the Thai baht (THB) against the Chinese yuan (CNY) during the Novel Coronavirus 2019 (COVID-19) pandemic by comparing a genetic algorithm (GA)-based subset autoregressive integrated moving average (ARIMA) model to the classical ARIMA model. Data was gathered from April 1, 2020 to April 14, 2022. Forecast accuracy was measured by mean absolute percentage error (MAPE), root mean squared error (RMSE) and mean absolute error (MAE). A GARI program was developed using non-seasonal time series prediction, with the best model ARIMA(4, 1, {1, 5}) forecasting daily CYN/THB exchange rate attaining a nadir of MAPE, RMSE and MAE at 1.2180%, 0.066674 and 0.064061, respectively. These findings indicate that the GA-based subset ARIMA model via GARI program outperformed the classical ARIMA model in the auto ARIMA with Python. This program may be applicable for predicting other foreign exchange rates and non-seasonal time series data.

**Keywords:** ARIMA model · Foreign exchange rates · Genetic algorithm · Time series · Chinese yuan currency

## 1 Introduction

Currently, each country has a different currency to use as a medium of exchange. In a conduct of international financial transactions, foreign exchange is required to be involved. In particular, those who need to study the trends in daily exchange rates for a month ahead including investors in foreign stocks, individuals who buy/sell foreign products, or related companies.

Foreign exchange rates are quantitative data that can be collected in many forms. One of the most popular formats is the time series, which is a sequence of discrete-time data collected chronologically in succession. Examples of time series are stock indices daily closing prices, wholesale weekly prices of agricultural products, and the monthly temperature average. We can say that the data correlated or recorded in conjunction with time is very interesting. This information can also reflect a particular event that occurred at that time. For instance,

- The baht was weakened sharply from 25 to 55 per *U.S. dollar (USD)* on July 2, 1997 as a result of speculation on the baht that has continued since the beginning of 1997. This affected lacking confidence of private businesses in the baht. It was deemed necessary to promulgate a floating exchange rate system by the Ministry of Finance and Bank of Thailand.
- The yuan was valued at 6.9999 and weaken to 7.0240 against the USD as of August 1, 2019. It was the first time that the yuan was dropped to 7 in more than a decade. As a result of trade war, China would let the yuan depreciate in order to help export sectors. Therefore, investors in global financial markets were concerned.
- On February 28, 2022, the ruble fell sharply to around 100.96 per USD compared to 83.5 as of February 23, 2022, which is the day before Russia's invasion of Ukraine.

It can be seen from the above three events that the big events happening in a country have impacted on the fluctuation of the country's currency. How could it be if we know the exchange rates in advance? Of course, it gives information to develop data-driven strategies and make decisions for international investors, a person or business which sells/buys goods from abroad, and those who are planning a trip oversea. These are the reason why we are interested in foreign exchange rate forecasting in the form of time series data in this paper.

Many researchers have applied statistical models and machine learning algorithms to time series forecasting, such as *Autoregressive Integrated Moving Average (ARIMA)*, *Nonlinear Autoregressive (NAR)* neural network, *Susceptible-Infectious-Recovered (SIR)*, *Long Short-Term Memory (LSTM)*, *genetic algorithm (GA)*, and *Artificial Neural Network (ANN)*, etc., which can be found in the following publications.

The best ARIMA model presented in [20] was selected by considering the smallest values of the criteria *Akaike Information Criterion (AIC)*, *Schwartz Information Criterion (SIC)*, *Mean Absolute Error (MAE)*, *Root Mean Squared Error (RMSE)*, and *Mean Absolute Percentage Error (MAPE)* in order to predict average daily share price indices of Square Pharmaceuticals Limited with non-stationary data series. Besides ARIMA, a commonly-used statistical model for time series forecasting is *Exponential Smoothing (ETS)*, which has automatic prediction strategies in Python package called auto ETS. Both ARIMA and ETS were applied in [16] to estimate daily exchange rates of the Romanian Leu against other currencies.

Swaraj et al. proposed in [23] a new model ARIMA-NAR that combines the ARIMA model with the NAR algorithm for prediction of COVID-19 cases

in India. The ARIMA-NAR model provided better prediction results with low values of RMSE, MAE, and MAPE compared to the single ARIMA model. This article claimed that the ARIMA-NAR model outperforms the SIR and LSTM algorithms for short-term forecasts. Moreover, the study [17] verified inability of the SIR in the long term forecast through the outbreak COVID-19 datasets in Isfahan province of Iran. Recently, the neural network LSTM has been tested its efficacy via streamflow prediction at ten river gauge stations across various climatic regions of the western United States [11], and power consumption in some French cities [15].

A two-level multi-objective GA was established by Al-Douri et al. in [1] to optimize the prediction of time series data on fans used in road tunnels according to the data from the Swedish Transport Administration. The two levels consist of a multi-objective GA implementation and the multi-objective GA utilization to identify an appropriate forecasting. The forecasting data for two life cycle costs obtained from their proposed models was neither realistic nor close to the actual data. A forecasting method integrating GA and *Autoregressive Moving Average (ARMA)*, or briefly called GA-ARMA, was proposed by Ervural et al., [7] to predict the monthly natural gas consumption of İstanbul collected from January 2004 to October 2015, where the fitness function of the GA was specified by MAPE. In [19], the GA was also used to identify ARMA, ARIMA and SARIMA models applied to semiconductor industry in terms of DRAM price forecasting. In addition, [25] has recently utilized a hybrid model of the GA and ARIMA, or briefly called GA-ARMA, to predict drought in synoptic station of Tabriz in northwestern Iran by investigating the Standard Precipitation and Evapotranspiration Index in the short-term, mid-term, and long-term steps of 53 years period. By comparing with the traditional models, the articles [7, 19] and [25] summarized that the GA-based models provided more accurate results.

Foreign exchange rate forecasting has been studied in the literature [3, 13, 24] and [25]. The yearly exchange rates of Kazakhstan currency: USD/KZT, EUR/KZT and SGD/KZT over the period from 2006 to 2014 were analyzed in [25] using MAE, MAPE and RMSE to measure the forecast accuracy. [13] developed three ANN based forecasting models using standard backpropagation, scaled conjugate gradient, and Bayesian regression to predict currency exchange rates of Australian dollar with six other currencies: USD, GBP, JPY, SGD, NZD and CHF. All the ANNs outperformed the traditional ARIMA model. Furthermore, forecasting exchange rate between Thai baht and U.S. dollar was investigated using time series analysis found in [3], and using data mining technique found in [24]. In 2014, Bowornchockchai [3] used Box-Jenkins and Holt's methods to forecast the monthly exchange rate. The author reported that the Box-Jenkins is the most suitable one. In 2020, the nine algorithms: Naive Bayes, generalized linear model, logistic regression, fast large margin, deep learning, decision tree, random forest, gradient boosted trees, and support vector machine were applied in [24] to predict the monthly exchange rate. The results of the study showed that logistic regression was reached the highest accuracy together with the three most significant correlated factors: U.S. dollar price index, gold price, and Nas-

daq price index. In recent works, the LSTM neural network has been used to forecast foreign exchange rates in [4,21] and [27], or even directional movement of Forex data in [28].

The remainder of the paper is organized as follows. In the next section, we define ARIMA model mathematically together with introducing auto ARIMA at the end of the section. The concept of subset ARIMA model based on GA is explained in Sect. 3. We add user's guide for using our developed program GARI as well as a link for readers who want to download and install the program. The performance of the model presented in Sect. 4 is tested using the criteria MAPE, RMSE and MAE, where the lowest value of them gives the highest accuracy model. The conclusion of the article is addressed in the last section.

## 2 ARIMA Model

Time series data is a data set collected sequentially in succession over equal periods of time. The period can be considered as a day, a month, a quarter, or a year. However, time series analysis is necessary to take into account the following four components of variation: trend, seasonal, cyclical and irregular components. The goal of the analysis is to identify relationships between observed values in the past through a model and use it to predict future values. The traditional forecasting models ARMA and ARIMA are described as follows.

The ARMA (Autoregressive Moving Average) model was popularized in 1970 by Box and Jenkins [9] developed from the Box-Jenkins method. The model is a combination of two models between *Autoregressive (AR)* and *Moving Average (MA)*, which are introduced in Definitions 2 and 3, respectively. Since ARMA is the most effective linear model for stationary time series forecasting compared with the pure AR and MA models, it is a popular and widely used model nowadays. The concept of the stationary time series can be written as the mathematical definition below.

**Definition 1.** Let  $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$  be the set of indices. The time series  $\{X_t : t \in \mathbb{Z}\}$  is said to be **stationary** if the following three statements hold:

1.  $E(X_t) = \mu$  for all  $t \in \mathbb{Z}$ .
2.  $\text{Var}(X_t) = \sigma^2 < \infty$  for all  $t \in \mathbb{Z}$ .
3.  $\text{Cov}(X_r, X_s) = \text{Cov}(X_{r+t}, X_{s+t})$  for all  $r, s, t \in \mathbb{Z}$ .

In other words, the time series  $\{X_t : t \in \mathbb{Z}\}$  is stationary if the mean and variance values are constants without depending on time  $t$ , while the covariance depends on  $r$  and  $s$  only through their difference  $|r - s|$ .

**Definition 2.** Let  $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$  be the set of indices. The **AR model** of order  $p$ , denoted by  $\text{AR}(p)$ , is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \epsilon_t \quad (1)$$

Here,  $\{X_t : t \in \mathbb{Z}\}$  is a stationary time series where  $\phi_1, \phi_2, \dots, \phi_p$  are parameters of the model such that  $\phi_p \neq 0$  and  $\epsilon_t$  is a white noise error at time  $t$ .

Let  $t \in \mathbb{Z}$ . When the mean  $\mu$  of  $X_t$  is nonzero, we can replace  $X_{t-i}$  of Eq. (1) with  $X_{t-i} - \mu$  for each  $i \in \{0, 1, \dots, p\}$  and obtain

$$\begin{aligned}
 X_t - \mu &= \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \dots + \phi_p(X_{t-p} - \mu) + \epsilon_t \\
 X_t &= (\mu - \mu\phi_1 - \mu\phi_2 - \dots - \mu\phi_p) + (\phi_1X_{t-1} + \phi_2X_{t-2} + \dots + \phi_pX_{t-p}) + \epsilon_t \\
 X_t &= \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p) + (\phi_1X_{t-1} + \phi_2X_{t-2} + \dots + \phi_pX_{t-p}) + \epsilon_t \\
 X_t &= \beta + \phi_1X_{t-1} + \phi_2X_{t-2} + \dots + \phi_pX_{t-p} + \epsilon_t,
 \end{aligned} \tag{2}$$

where  $\beta = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p)$ . It can be inferred that Eq. (2) is similar to the regression model, which has independent variables as its own previous values. Therefore, we call Eq. (1) an autoregression model as a result of ‘‘auto’’ here referring to ‘‘self’’ in this context.

**Definition 3.** The **MA model** of order  $q$ , denoted by  $MA(q)$ , is of the form

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \dots + \theta_q\epsilon_{t-q}, \tag{3}$$

where  $\theta_1, \theta_2, \dots, \theta_q$  are parameters of the model such that  $\theta_q \neq 0$  and  $\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q}$  are the current and various past values of a white noise error.

The AR and MA models lead to the concept of ARMA model, which was firstly described in 1951 by Whittle [26] as presented in the definition below.

**Definition 4.** Let  $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$  be the set of indices. The **ARMA model** of a given stationary time series  $\{X_t : t \in \mathbb{Z}\}$  with  $p$  autoregressive terms and  $q$  moving average terms, denoted by  $ARMA(p, q)$ , is of the form

$$X_t = \phi_1X_{t-1} + \phi_2X_{t-2} + \dots + \phi_pX_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \dots + \theta_q\epsilon_{t-q}, \tag{4}$$

where  $\phi_1, \phi_2, \dots, \phi_p$  and  $\theta_1, \theta_2, \dots, \theta_q$  are parameters of the model such that  $\phi_p \neq 0$  and  $\theta_q \neq 0$  together with the current and various past values of a white noise error  $\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q}$ , respectively.

It is easy to see from Eq. (4) that the ARMA model is a generalization of AR and MA models. That is,

- if  $q = 0$ , then the ARMA model becomes autoregressive model of order  $p$ , or  $AR(p)$ ,
- if  $p = 0$ , then the ARMA model turns into moving average model of order  $q$ , or  $MA(q)$ .

In practice, the ARMA model cannot be applied to ‘‘non-stationary’’ time series data but it is necessary to transform the data to be ‘‘stationary’’ first by performing difference operation. This is the underlying process of **ARIMA model**:

Let  $\{X_t : t \in \mathbb{Z}\}$  be the original time series with non-stationary data.

$$\text{If } \Delta^d X_t \text{ is } ARMA(p, q), \text{ then } X_t \text{ is } ARIMA(p, d, q),$$

The term  $\Delta^d X_t$  is called **differencing operator** of order  $d$ , which comes from “Integrated” concept. Throughout the paper, the notation ARIMA( $p, d, q$ ) refers to the non-seasonal ARIMA models with prediction equation:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}, \quad (5)$$

where  $Z_* = \Delta^d X_*$  for all indices  $* \in \{t, t-1, \dots, t-p\}$ . The nonnegative integers  $p, d$  and  $q$  represent as the table below.

Value	Meaning
$p$	The number of the AR terms
$d$	The number of differences needed for stationarity
$q$	The number of lagged forecast errors in the prediction equation

The values  $p, d$  and  $q$  can be generated in auto ARIMA with python using “pmdarima” package. The steps how it works is explained as seen in Algorithm 1. The model selection criteria AIC and BIC showing up in the algorithm are described in [8].

---

**Algorithm 1** : The auto ARIMA procedure.

---

- Step 1: Use *augmented Dickey-Fuller (ADF)* test to check whether the time series data is stationary or not.
    - If the data is stationary, set  $d = 0$ .
    - If the data is non-stationary, perform the difference to find the order  $d$  of the model.
  - Step 2: Identify the suitable orders  $p$  and  $q$  of the corresponding AR and MA parts by using plots of the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function), respectively.
  - Step 3: Estimate the parameters  $\phi_1, \phi_2, \dots, \phi_p$  and  $\theta_1, \theta_2, \dots, \theta_q$  of the model.
  - Step 4: Improve the model to be the best fit one by comparing AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values.
- 

### 3 GA-Based Subset ARIMA Model

GA (Genetic Algorithm) was invented in early 1970s by Holland and his students at the University of Michigan as an evolutionary optimization method developed from a random search algorithm in conjunction with the survival concept of the strongest individuals according to Darwin’s theory of evolution. Due to the ability of GA to solve complex and nonlinear problems, it has been applied to the field of artificial intelligence in recent times.

The GA searches for the globally accepted optimal solution, which is simply called the best solution, to an optimization problem within a reasonable time.

Initially, the algorithm generates randomly a population of candidate solutions, each of which is treated as a chromosome (an individual) of the population. Later, fitness value for each chromosome is evaluated. The chromosomes selected based on their fitness values in selection process become parents in reproduction process in order to produce offspring of the next generation by performing evolutionary strategies including crossover and mutation.

Recalling Algorithm 1, let  $d$  be a fixed order of differencing according to Step 1. In Step 3, we can apply a library named “statsmodels” in Python to estimate the parameters. The idea of GA-ARIMA arises from the use of GA to search for the best ARIMA model reached the highest prediction accuracy instead of performing traditional Steps 2 and 4. Since ARIMA with low orders can sufficiently model most time series. To restrict the search space of the GA, the orders  $p$  and  $q$  focused here do not exceed 5, i.e.,  $p_{\max} = 5$  and  $q_{\max} = 5$ , where  $p_{\max}$  and  $q_{\max}$  denote the maximum possible orders corresponding to MA and AR models. With those orders, the prediction Eq. (5) is expressed as

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \phi_4 Z_{t-4} + \phi_5 Z_{t-5} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \theta_4 \epsilon_{t-4} + \theta_5 \epsilon_{t-5}, \tag{6}$$

where  $Z_* = \Delta^d X_*$  for all indices  $* \in \{t, t - 1, \dots, t - 5\}$ . However, if some coefficient parameters of Eq. (6) are set to zero, the model is called a **subset ARIMA model**, which is an extended version of the ARIMA model introduced by Lee and Fambro [14]. Throughout the paper, the subset ARIMA model is denoted by  $\text{ARIMA}(P, d, Q)$ , where  $P$  and  $Q$  are nonempty subset of  $\{1, 2, 3, 4, 5\}$ . For instance, a model  $\text{ARIMA}(\{1, 2, 3, 4\}, 1, \{1, 5\})$  implies that the parameters  $\phi_1, \phi_2, \phi_3, \phi_4$  exist in the AR part, and  $\theta_1, \theta_5$  exist in the MA part, and the others are set to zero. This leads to the prediction equation

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \phi_4 Z_{t-4} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_5 \epsilon_{t-5}, \tag{7}$$

where  $Z_* = \Delta X_*$  for all indices  $* \in \{t, t - 1, t - 2, t - 3, t - 4\}$ . Note that if the set  $P$  (or  $Q$ ) is  $\{1, 2, \dots, r\}$ , we represent it as the number at the end  $r$  for convenience. So, the above model is briefly written by  $\text{ARIMA}(4, 1, \{1, 5\})$ . To generate the possible sets  $P$  and  $Q$  of the subset ARIMA models and search for the best one using the GA, we need to define a chromosome representation used in the algorithm as a string of length  $p_{\max} + q_{\max} = 5 + 5 = 10$  depicted in Fig. 1. For each  $i, j \in \{1, 2, 3, 4, 5\}$ , the  $u_i$  and  $v_j$  on the chromosome are valued as (8), whose values indicate the existence of the  $\phi_i$  and  $\theta_j$  in the AR and MA parts of the model, respectively.

$$u_i = \begin{cases} 1, & \text{if } i \in P; \\ 0, & \text{if } i \notin P, \end{cases} \text{ and } v_j = \begin{cases} 1, & \text{if } j \in Q; \\ 0, & \text{if } j \notin Q. \end{cases} \tag{8}$$

$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

**Fig. 1.** a chromosome representation of the GA in GA-ARIMA model.



With the following chromosome for example,

1	0	0	1	1	1	1	0	0	0
---	---	---	---	---	---	---	---	---	---

we can convert it to the model ARIMA( $\{1, 4, 5\}, d, \{1, 2\}$ ), where  $d$  is an appropriate differencing order according to the ADF test, whose prediction equation is

$$Z_t = \phi_1 Z_{t-1} + \phi_4 Z_{t-4} + \phi_5 Z_{t-5} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2},$$

where  $Z_* = \Delta^d X_*$  for all indices  $* \in \{t, t - 1, t - 4, t - 5\}$ .

For the improvement step of the model, we try to find out the best chromosome with the smallest value of criteria MAPE, RMSE and MAE as our error measures. Let  $n$  be the total number of fitted points. The above criteria can be calculated by

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{X_t - \hat{X}_t}{X_t} \right|, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (X_t - \hat{X}_t)^2}, \tag{10}$$

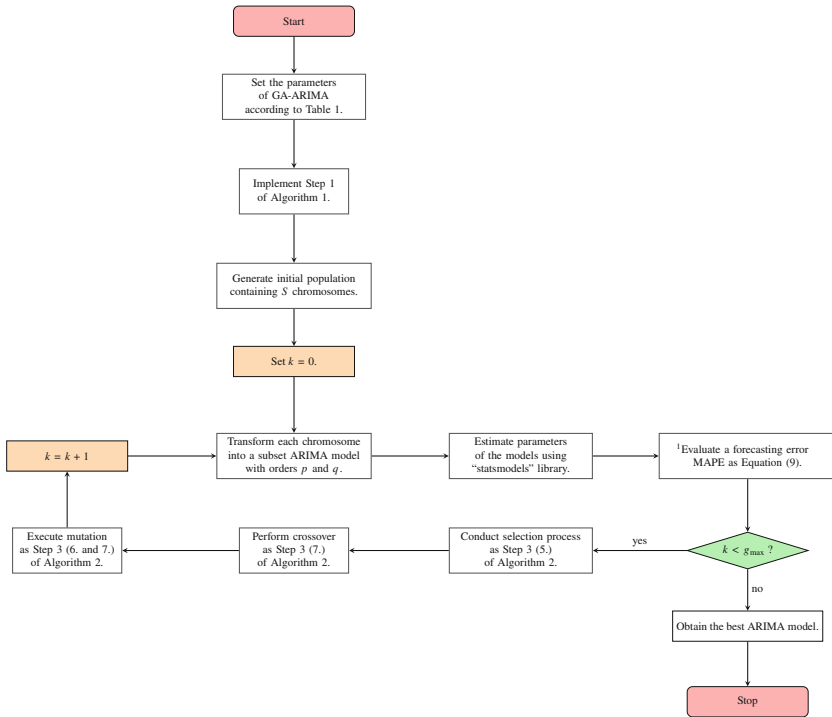
$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |X_t - \hat{X}_t|, \tag{11}$$

where  $X_t$  and  $\hat{X}_t$  are the actual and forecast values at time  $t$ , respectively.

As the preceding, the proposed procedure in accordance with GA-based subset ARIMA model can be concluded as a flowchart in Fig. 2. In this work, we also develop an application program named ‘‘GARI’’ that comes from GA-ARIMA for Investment in terms of future trends prediction. This program is suitable for not only predicting the exchange rate data but applying to any time series data with non-seasonality and univariate (stationarity) non-stationarity that could be made stationarity by differencing. Note that there are a number of non-stationary time series data, which does not work for taking over-differencing in order to ensure the stationarity according to the paper [10]. Readers can download and install the program from the link: <https://drive.google.com/drive/folders/1Z9vEpzfqICDGdL8bjyEV-fUL7ZLsNn5e?usp=sharing>. When the installation is complete, the main window will appear displayed as Fig. 3. The parameters of the program defined as Table 1 need to be set at the beginning. Users can learn more about their description in Step 3 of Algorithm 2. In that algorithm, we record the step-by-step implementation of GARI. Moreover, our program provides two additional options:

1. FAST GA: This option enables the users to forecast more comfortably, and the maximum acceptable error is required. Without this option, please choose “disable” mode.
2. Show detail: The chromosomes in each generation will be displayed in details on the lower left-hand side of GARI’s window. With choosing “disable’ mode, the details will be concise.

Normally, we would be able to select “disable” mode for the above options. The program will lead us to the best ARIMA model displayed in “Summary” window. In addition, time series plots of the exchange rate including actual and forecast data are illustrated in “Graph” window. Eventually, we can see the forecast values in “Table” window.



**Fig. 2.** Flowchart for the GA-ARIMA procedure. (<sup>1</sup> The evaluation could be replaced by RMSE and MAE.)

**Table 1.** Notation and meaning of parameters used in the algorithm.

Symbol	Meaning	Setting value
$C$	Confidence level (in percentage).	95%
$T$	The amount of testing data.	30 days
$g_{\max}$	The maximum number of generations in the algorithm, written herein as “Max generation” for short	200
$S$	Population size.	10
$s$	The number of survivors	5
$m$	The number of mutation points	1
$R_m$	Mutation rate (in percentage).	30%
$R_c$	Crossover rate (in percentage).	80%

**Algorithm 2 :** The GA-ARIMA procedure to predict CNY/THB via our proposed program.

- 
- Step 1: Upload a data set accepted only file types ‘.csv’ and ‘.xlsx’.
- Step 2: Choose a column name addressed the daily exchange rate data.
- Step 3: Set inputs related to predicting options and genetic algorithm setting mentioned in Table 1. Explanation of the inputs is presented as below.
1. The accuracy of the model is represented as an interval at  $C = (1-\alpha)100\%$  confidence level, whose value is 90%, 95% and 99% typically corresponding to its significance level  $\alpha$  as 0.10, 0.05 and 0.01, respectively.
  2. The  $T$  can be specified by either percentage as  $T\%$  or the number of data points, which directly means that the last  $T$  points become the testing data. The number of data in total, training and testing will be shown when clicking the button “Show detail”.
  3. The value of  $g_{\max}$  is used in this work as a stopping criterion of the GA.
  4. The population size  $S$  represents the number of chromosomes in each generation.
  5. In selection process, the first  $s$  chromosomes ranked from the best fitness value to the worst one can survive to the next generation This means a number of chromosomes will be randomly generated to fulfill the population size at that generation.
  6. In mutation, since our chromosome representation is a binary string of genes, we randomly select  $m$  genes and flip their values from 1s to 0s and vice versa.
  7. The rates  $R_c$  and  $R_m$  are nonnegative integers in the range  $[0, 100]$ . The survey [2] informed that the crossover rate  $R_c$  is typically valued in the range  $[60, 100]$ . While, the appropriate value of  $R_m$  for a given optimization problem is an open research issue.
    - For each parents, we generate a random integer  $R$  between 0 and 100 and perform a single point crossover to the parents that produce offspring if  $R < R_c$ . Otherwise, such execution is none.
    - For each chromosome, we generate  $R$  similarly as above. If  $R < R_m$ , the chromosome is mutated by flipping on its genes according to the number of mutation points  $m$ . Otherwise, such execution is none.
- Step 4: Choose either one of the forecast accuracy measures MAPE, RMSE or MAE, whose descriptions can be seen in Table 1 of [5] and on pages 5–6 of [12]. These computations are written as Eqs. (9)–(11).
- Step 5: Click the button “Start”. The forecasting results will be displayed at GARI’s windows in a while. The running time depends on the size of input data and the  $g_{\max}$  setting. However, the users can stop running anytime by clicking button “Stop”.
-

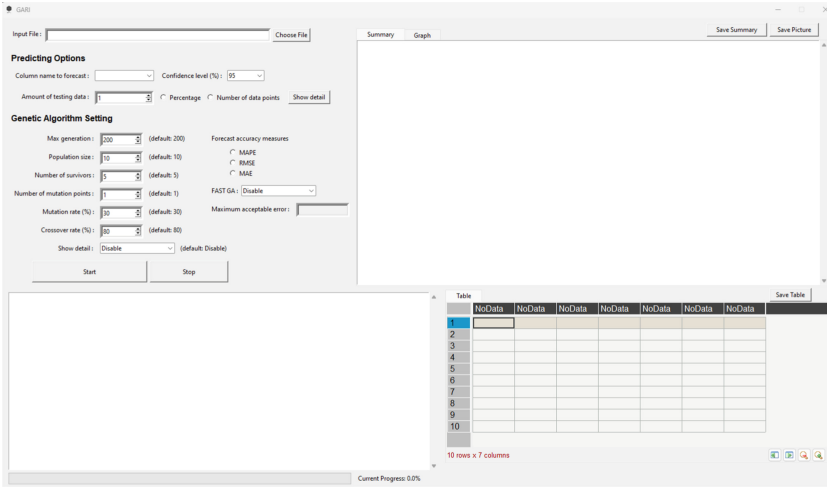


Fig. 3. Main window of GARI.

## 4 Results

The currency exchange rate between *Thai Baht (THB)* and *Chinese Yuan (CNY)* was collected from Bank of Thailand during 2020–2022 in two data sets:

1. The daily exchange rate from April, 1, 2021 to April, 14, 2022 (one year period) without the weekend consists of overall 271 observations.
2. The daily exchange rate from April, 1, 2020 to April, 14, 2022 (two years period) without the weekend consists of overall 532 observations.

The last 30 days of those data sets, i.e., March, 4, 2022 – April, 14, 2022, are used for testing validity of the models. By using auto ARIMA in Python, the best model for the first data set is ARIMA(1, 1, 2). For the second one, it gives the model ARIMA(1, 1, 1). The performance measurement of the ARIMA models obtained from auto ARIMA and the subset ARIMA models obtained from our program GARI can be seen in Table 2. This table reports that the GARI provides less values of MAPE, MAE and RMSE compared to auto ARIMA.

For the two years period, we accomplish the same model ARIMA(4, 1, {1, 5}) based on those error measures. Its prediction equation as shown in Table 2 could be transformed to the one in terms of the original time series  $\{X_t : t \in \mathbb{Z}\}$  as

$$\begin{aligned}
 X_t - X_{t-1} &= 0.7826(X_{t-1} - X_{t-2}) + 0.0480(X_{t-2} - X_{t-3}) - 0.0105(X_{t-3} - X_{t-4}) \\
 &\quad + 0.1752(X_{t-4} - X_{t-5}) + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5} \\
 X_t &= (0.7826 + 1)X_{t-1} + (0.0480 - 0.7826)X_{t-2} + (-0.0105 - 0.0480)X_{t-3} \\
 &\quad + (0.1752 + 0.0105)X_{t-4} - 0.1752X_{t-5} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5},
 \end{aligned}$$

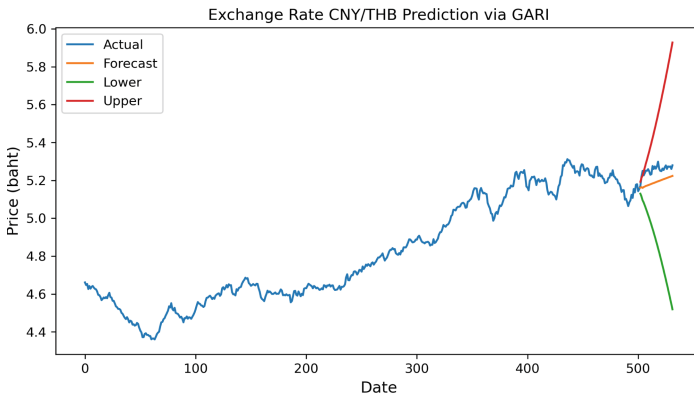
**Table 2.** Forecasting results from auto ARIMA and GARI for the daily CYN/THB exchange rate in two periods of data collection.

Data collection period	Criteria	Auto ARIMA	GARI	The best model from GARI
1 Year	MAPE	1.8381%	1.6937%	ARIMA(5, 1, {2, 4, 5}) has the prediction equation $Z_t = 0.1055Z_{t-1} - 0.3152Z_{t-2} + 0.0371Z_{t-3}$ $+ 0.2075Z_{t-4} + 0.4990Z_{t-5} + \epsilon_t$ $+ 0.4526\epsilon_{t-2} - 0.0124\epsilon_{t-4} - 0.6550\epsilon_{t-5}$ .
	MAE	0.096705	0.089106	
	RMSE	0.099473	0.091623	
2 Years	MAPE	1.9276%	1.2180%	ARIMA(4, 1, {1, 5}) has the prediction equation $Z_t = 0.7826Z_{t-1} + 0.0480Z_{t-2} - 0.0105Z_{t-3}$ $+ 0.1752Z_{t-4} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5}$ .
	MAE	0.101412	0.064061	
	RMSE	0.104062	0.066674	

which implies

$$\hat{X}_t = 1.7826X_{t-1} - 0.7346X_{t-2} - 0.0585X_{t-3} + 0.1857X_{t-4} - 0.1752X_{t-5} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5}. \tag{12}$$

Furthermore, the illustration in Fig.4 shows the time series plot of the CNY/THB exchange rate data. Within the scope of prediction, the values are clarified in Table 3. We observe that the forecast values lie on the range of 95% confidence interval from March, 9, 2022 onwards.



**Fig. 4.** Actual and forecast values of the daily CYN/THB exchange rate collected in two years.

At the end of this section, we additionally record the errors from ETS model (via auto ETS in Python) and LSTM algorithm in Table 4. The GARI can

**Table 3.** Actual and forecast values with its lower and upper values at 95% confidence level obtained by the best model ARIMA(4, 1, {1, 5}) with its forecasting equation as (12) for the daily CYN/THB exchange rate collected in two years.

Date	Actual	Forecast	95% confidence interval	
			Lower	Upper
4-Mar-22	5.1716	5.161101256	5.129606165	5.192596346
7-Mar-22	5.2197	5.164041509	5.114992439	5.21309058
8-Mar-22	5.2503	5.160236278	5.09441576	5.226056796
9-Mar-22	5.2263	5.164947343	5.08302293	5.246871756
10-Mar-22	5.2384	5.167678293	5.06709499	5.268261597
11-Mar-22	5.2518	5.170596575	5.051533326	5.289659825
14-Mar-22	5.2506	5.172295425	5.034476495	5.310114356
15-Mar-22	5.2594	5.174561539	5.017711585	5.331411492
16-Mar-22	5.2475	5.176864204	5.000304475	5.353423932
17-Mar-22	5.2299	5.179268321	4.982419653	5.37611699
18-Mar-22	5.233	5.18153403	4.963841462	5.399226598
21-Mar-22	5.2762	5.183795279	4.944746162	5.422844396
22-Mar-22	5.26	5.18605173	4.925113781	5.446989679
23-Mar-22	5.274	5.188323436	4.904970204	5.471676668
24-Mar-22	5.2618	5.190582652	4.88429728	5.496868023
25-Mar-22	5.2744	5.192832095	4.863113847	5.522550342
28-Mar-22	5.299	5.19507229	4.841431219	5.548713361
29-Mar-22	5.2605	5.197307582	4.81926417	5.575350994
30-Mar-22	5.2502	5.199536508	4.796621393	5.602451623
31-Mar-22	5.2486	5.201758602	4.773513265	5.630003939
1-Apr-22	5.2627	5.203973475	4.749949807	5.657997143
4-Apr-22	5.2548	5.206181578	4.725941451	5.686421705
5-Apr-22	5.2689	5.208382992	4.701497816	5.715268168
6-Apr-22	5.2791	5.210577727	4.676628179	5.744527275
7-Apr-22	5.2605	5.212765718	4.651341437	5.77419
8-Apr-22	5.274	5.214946998	4.62564627	5.804247725
11-Apr-22	5.2714	5.217121598	4.599551037	5.83469216
12-Apr-22	5.2741	5.219289551	4.573063796	5.865515306
13-Apr-22	5.2602	5.22145087	4.546192309	5.896709432
14 Apr-22	5.2796	5.223605574	4.518944071	5.928267077

achieve more accurate than the ETS model but cannot achieve when comparing to LSTM. Since the GARI is developed on the basis of the traditional model ARIMA, it is not possible to reach superior to deep leaning-based algorithm LSTM, which was also confirmed in the articles [6, 18] and [22].

**Table 4.** The MAPEs, MAEs and RMSEs from GARI compared to auto ETS and LSTM for the daily CYN/THB exchange rate in two periods of data collection.

Data collection period	Criteria	GARI	Auto ETS	LSTM
1 Year	MAPE	1.6937%	1.9240%	0.7309%
	MAE	0.089106	0.101218	0.038439
	RMSE	0.091623	0.103869	0.042586
2 Years	MAPE	1.2180%	1.8850%	0.6097%
	MAE	0.064061	0.099165	0.032071
	RMSE	0.066674	0.101783	0.037495

## 5 Conclusion

This paper utilized GA-based subset ARIMA model with combining best sides of genetic algorithm and ARIMA model to improve prediction accuracy of the single ARIMA model. The ARIMA is a popular model to analyze stationary and non-stationary univariate time series data. The use of the genetic algorithm is to determine which ARIMA model is the best. In addition, we developed a program named GARI to predict time series data on the basis of GA-ARIMA procedure. Our program was applied to forecast the daily exchange rate for the Thai baht against the Chinese yuan reached higher forecast accuracy compared to using auto ARIMA with Python based on the statistical ARIMA model.

## References

1. Al-Douri, Y.K., Hamodi, H., Lundberg, J.: Time series forecasting using a two-level multi-objective genetic algorithm: a case study of maintenance cost data for tunnel fans. *Algorithms* **11**(8), 123 (2018)
2. Boussaïd, I., Lepagnot, J., Siarry, P.: A survey on optimization metaheuristics. *Inf. Sci.* **237**, 82–117 (2013)
3. Bowornchockchai, K.: Forecasting exchange rate between Thai Baht and the US dollar using time series analysis. *Int. J. Math. Comput. Sci.* **8**(8), 1186–1191 (2016)
4. Dautel, A.J., Härdle, W.K., Lessmann, S., Seow, H.V.: Forex exchange rate forecasting using deep recurrent neural networks. *Digit. Financ.* **2**(1), 69–96 (2020)
5. Dzikevičius, A., Šaranda, S.: Smoothing techniques for market fluctuation signals. *Bus. Theory Pract.* **12**(1), 63–74 (2011)
6. Elsaraiti, M., Merabet, A.: A comparative analysis of the ARIMA and LSTM predictive models and their effectiveness for predicting wind speed. *Energies* **14**(20), 6782 (2021)
7. Ervural, B.C., Beyca, O.F., Zaim, S.: Model estimation of ARMA using genetic algorithms: a case study of forecasting natural gas consumption. *Procedia Soc. Behav. Sci.* **235**, 537–545 (2016)
8. Fabozzi, F.J., Focardi, S.M., Rachev, S.T., Arshanapalli, B.G.: *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*. Wiley, Hoboken (2014)

9. George, E., Jenkins, G.M., Reinsel, G.C.: *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken (1970)
10. Hossain, Z., Rahman, A., Hossain, M., Karami, J.H.: Over-differencing and forecasting with non-stationary time series data. *Dhaka Univ. J. Sci.* **67**(1), 21–26 (2019)
11. Hunt, K.M., Matthews, G.R., Pappenberger, F., Prudhomme, C.: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western united states. *Hydrol. Earth Syst. Sci. Discuss.* **26**, 5449–5472 (2022)
12. Jierula, A., Wang, S., Oh, T.M., Wang, P.: Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Appl. Sci.* **11**(5), 2314 (2021)
13. Kamruzzaman, J., Sarker, R.A.: Forecasting of currency exchange rates using ANN: a case study. In: *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, vol. 1, pp. 793–797. IEEE (2003)
14. Lee, S., Fambro, D.B.: Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transp. Res. Rec.* **1678**(1), 179–188 (1999)
15. Mahjoub, S., Chrifi-Alaoui, L., Marhic, B., Delahoche, L.: Predicting energy consumption using LSTM, multi-layer GRU and drop-GRU neural networks. *Sensors* **22**(11), 4062 (2022)
16. Maria, F.C., Eva, D.: Exchange-rates forecasting: exponential smoothing techniques and ARIMA models. *Ann. Faculty Econ.* **1**(1), 499–508 (2011)
17. Moein, S., et al.: Inefficiency of sir models in forecasting Covid-19 epidemic: a case study of Isfahan. *Sci. Rep.* **11**(1), 1–9 (2021)
18. Muncharaz, J.O.: Comparing classic time series models and the LSTM recurrent neural network: an application to s&p 500 stocks. *Financ. Markets Valuation* **6**(2), 137–148 (2020)
19. Ong, C.S., Huang, J.J., Tzeng, G.H.: Model identification of ARIMA family using genetic algorithms. *Appl. Math. Comput.* **164**(3), 885–912 (2005)
20. Paul, J.C., Hoque, M.S., Rahman, M.M.: Selection of best ARIMA model for forecasting average daily share price index of pharmaceutical companies in Bangladesh: a case study on Square Pharmaceutical Ltd. *Glob. J. Manag. Bus. Res.* **13**, 14–25 (2013)
21. Qu, Y., Zhao, X.: Application of LSTM neural network in forecasting foreign exchange price. In: *Journal of Physics: Conference Series*, vol. 1237, p. 042036. IOP Publishing (2019)
22. Siami-Namini, S., Tavakoli, N., Namin, A.S.: A comparison of ARIMA and LSTM in forecasting time series. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401. IEEE (2018)
23. Swaraj, A., Verma, K., Kaur, A., Singh, G., Kumar, A., de Sales, L.M.: Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. *J. Biomed. Inform.* **121**, 103887 (2021)
24. Tepdang, S., Ponprasert, R.: Forecast of changes in exchange rate between Thai Baht and US dollar using data mining technique. *SNRU J. Sci. Technol.* **12**(3), 213–221 (2020)
25. Tlegenova, D.: Forecasting exchange rates using time series analysis: the sample of the currency of Kazakhstan. arXiv preprint [arXiv:1508.07534](https://arxiv.org/abs/1508.07534) (2015)
26. Whittle, P.: *Hypothesis Testing in Time Series Analysis*, vol. 4. Almqvist & Wiksells boktr. (1951)



27. Wijesinghe, S.: Time series forecasting: analysis of LSTM neural networks to predict exchange rates of currencies. *Instrumentation* **7**(4), 25 (2020)
28. Yıldırım, D.C., Toroslu, I.H., Fiore, U.: Forecasting directional movement of forex data using LSTM with technical and macroeconomic indicators. *Financ. Innov.* **7**(1), 1–36 (2021)



# Impacts of Countermeasure Program on the Covid-19 Pandemic in Asian Countries

Worrawat Saijai and Sukrit Thongkairat<sup>(✉)</sup>

Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University,  
Chiang Mai 50200, Thailand  
sukrit415@gmail.com

**Abstract.** The spread of the COVID-19 pandemic in 2020 has produced a large impact on various economic sectors. In this paper, we analyze the policies that were implemented during the epidemic using weekly data between March 16, 2020 and August 30, 2021 from over 70 economies in Asia (all countries in Asia and all provinces in China), with the observation number is 5,865 observations. The policies included in this investigation on their efficacy are those declared as School and Workplace Closures, Cancellation of Public Events and Gatherings, Stay-at-Home Restrictions, Face Coverings, Public Information Campaigns, International and Domestic Travel, Testing and Contact Tracing, Vaccination Policy, etc. We employed various models to find the best estimates and found out that the Poisson lasso model outperforms the others. The Poisson lasso model is remarkably helpful for variable selection as it shrinks the parameter values of unexplainable variables into zero. The results show some policies can decrease the total cases after one month implementation including the vaccination policies of the universal availability policy for vaccination and testing anyone showing COVID-19 symptoms, public information campaigns, international support. Some policy measures, however, were observed to be slow in decreasing the total number of cases such as the prohibition of mass gatherings and public events, international travel controls, closure of some business sectors or work categories, recommended closing or significantly reducing volume/route/means of transport available (especially public transportation), and not to leave home unless it is extremely necessary (stay-at-home requirements).

**Keywords:** Lasso regression · Covid-19 · Countermeasure programs

## 1 Introduction

An economic system is probably analogous to a cog to which a damage or problem that even very slightly occurs somewhere can adversely affect the normal operation of one or more or all parts of the economy. Also analogously, the pandemic of Coronavirus which has damaged the health of the entire world is not

only a health crisis but also an economic one throughout the globe. The pandemic crisis thus has posed a challenge for individuals, societies, and economies as well as providing a platform for testing policy instruments whether any of them can help the economic system to reduce, repair, and retain societies. The Coronavirus was first discovered in Wuhan, China in December 2019. Characteristically, the virus can rapidly spread and directly affect human respiratory system, causing pneumonitis. Since the spread can be transmitted from person to others through breathing, keeping the virus to spread only in the quarantine area is difficult (Ozili and Arun, 2020). Then, World Health Organization (WHO) has announced COVID-19 as a global pandemic on March 11, 2020 (Cucinotta and Vanelli, 2020).

The outbreak of COVID-19 clearly affected a country's economy, even the economically advanced countries are not an exception. The United Kingdom, for example, is one of the economically advanced countries having a good economic management system. With the arrival of this epidemic in the UK, some businesses had to close, and employers were made involuntarily inactive due to infection and panic. The mortality rate from COVID-19 in the UK was reportedly more than five times higher than other European countries (Hopkins, 2020; Monbiot, 2020), indicating a legacy of policy failures. Thus, the effects of various policies on the number of infection and death during the epidemic were advised to be examined to determine which policies should be strongly implemented.

However, the policies should be verified for two main reasons. Firstly, the such a crisis has never occurred in the world. So, the emergency policies for reducing the number of infected cases got well responses. Secondly, we must have plans, policies, and visions for repairing and retaining the economy when the crisis ends. The measurement of effect from various policies to control the disastrous events during the period of COVID-19 pandemic has been a focus of attention in many pieces of literature. One of the interesting features of policy has been the question of "Which policies work well to reduce the number of infected cases by COVID-19?"

In this paper, we are interested in a linear model that can answer research problems. Therefore, the relevant literature on this subject area was reviewed. Hoerl and Kennard (1970) developed and applied the statistical properties of the ridge regression in a linear model. Later, Mansson and Shukur (2011) have improved the ridge regression by using a Poisson ridge regression estimator. This popular method is used for estimation and regression to calculate high correlation coefficients by finding the log-likelihood function in the maximum likelihood estimation (MLE). However, the approximation of the regression coefficient with the ridge model will result in all the coefficients approaching zero or shrinking, making the coefficient approximation to be more stable. However, it does lack the qualification to select independent variables in the model. Tibshirani (1996) proposed a method for analyzing the Poisson lasso regression in a linear model. This method is more than just involving regression coefficient estimation but also the selection of reasonable independent variables into the model. The estimation of coefficient parameter of the lasso regression also had some limitations in the

selection of independent variables into a model with consistent properties, and later, Park and Hastie (2007) developed and presented the statistical properties of the lasso regression analysis for the regression model. This is an extension of Fan and Li (2001) work by modifying consistency properties in the Poisson regression model. As a result, this method of estimation has been more efficient in selecting variables into the model while reducing the estimation bias, making it is better than the conventional lasso methods. We then call this an improved method or “adaptive Poisson lasso method”.

Therefore, in this paper; to answer the research question in the case of Thailand, it is necessary to investigate the policy measures that were implemented during the epidemic, such as School and Workplace Closures, Cancellation of Public Events and Gatherings, Stay-at-Home Restrictions, Face Coverings, Public Information Campaigns, International and Domestic Travel, Testing and Contact Tracing, Vaccination Policy, etc. We use the lasso model as it is very easy to consider. This model includes a relationship to regression, best variable selection, and the connections between the lasso coefficient estimates.

The rest of the paper is structured as follows. Section 2 discusses the methodology which is mainly about the Poisson lasso regression model. Section 3 shows the variation of economic policies in several countries and the data description. Section 4 criticizes some of the policies and shows the empirical results. Section 5 provides the conclusion and recommendations on policy to counteract a pandemic.

## 2 Methods and Procedures

### 2.1 Definitions

We intend to estimate the regression model below using total cases per million ( $Y$ ) as a proxy variable, assuming that the death rate is small. (the death rate around the globe is 2.07% on September 21, 2021, as provided by ourworldin-data.org)

$$\ln Y_t = \sum_{i=1}^n \alpha_{1,t-i} \ln Y_{t-i} + \sum_{i=1}^n \alpha_{1,t-i} \ln X_{1,t-i} + \dots + \sum_{i=1}^n \beta_{m,t-i} \ln X_{m,t-i} + \sum_{i=1}^n \gamma_{1,t-i} \text{ Policy }_{1,t-i} + \dots + \sum_{i=1}^n \gamma_{q,t-i} \text{ Policy }_{q,t-i} \tag{1}$$

Here,  $Y$  is modeled by the autoregressive process with lag  $n$  and other  $m$  numerical and  $q$  factor variables. In the numerical variables, we include the variables that might affect the number of total cases per million, such as positive rate, total vaccinations per hundred, people fully vaccinated per hundred, people vaccinated per hundred, total boosters per hundred, together with some policy mechanisms like accumulative fiscal measures and emergency investment in healthcare. The factor variables are mostly the government policies, including school closing, workplace closing, cancellation of public events, restrictions on mass gathering,

limited public transportation, stay-at-home requirements, restrictions on internal movement, international travel control, income support, debt contract relief, public information campaigns, testing policy, contact tracing, investment in vaccines, facial coverings, vaccination policy, and protection of elderly people. Those are the policy measures that could help us understand more about the relative effectiveness of each policy and be able to provide the empirical results and policy suggestions. Moreover, we also put some demographic variables like population density, median age, the proportion of aged 65 and above elderly, GDP per capita, extreme poverty, cardiovascular death rate, diabetes prevalence, and life expectancy to help understand more in terms of basic characteristics that could contribute to the increase of total cases.

However, the Poisson lasso and Poisson adaptive regression will shrink some coefficients into zero, indicating that the lag variables have no causal effect on the total cases., while the ridge regression will reduce the size of some coefficients to be close to zero, which means that the variables have less causal effect. Therefore, we will show only the variables with coefficients that are non-zero. The negative coefficient shows the positive effect on the total cases reduction, meaning that the corresponding variable or policy helps reduce the total cases. While the positive coefficient will increase the total cases number or reduce the positive effect in the earlier weeks.

## 2.2 Statistical Analysis

### 2.2.1 Poisson Regression Model

Poisson regression is a generalized linear model form of regression analysis. It is often used for modeling count data. Let  $y = (y_1, \dots, y_n)^T$  be an  $n \times n$  vector of dependent variable with the vector of mean value as in the following

$$\mu = (\mu_1, \dots, \mu_n)^T \tag{2}$$

The Probability Mass Function (p.m.f.) of the Poisson regression can be shown as follows

$$f(y_i; \mu_i, X_i) = \frac{\mu_i^{(y_i)} \exp(-\mu_i)}{y_i!}, i = 1, 2, 3, \dots, \text{ and } \mu. \tag{3}$$

where  $\mu_i = \mu(X_i, \beta) = \exp(X_i^T \beta)$

$X_i$  is a matrix of independent variable  $i$  and  $\beta$  is a vector of unknown regression coefficients, which can be estimated by using maximum likelihood approach. Then, the Poisson regression can be shown in the following form

$$Y_i = E[Y_i] + \varepsilon_i, 1, 2, 3, \dots, n$$

where the mean value of response variable  $i$  is equal to  $\mu_i$

$$Y_i = \exp(X_i^T \beta) + \varepsilon_i$$

### 2.2.2 Poisson Lasso Regression

Park and Hastie (2007) invented Poisson Lasso regression by adding L1 regularization path into the generalized linear models. This model is useful and popular for coefficient estimation and variable selection, enabling us to reach both ends at the same time. The estimated coefficients can be found by using Log Likelihood function under L1 penalty on  $\beta$ ,  $L_1 = \|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ . Then, the definition of  $\hat{\beta}_{\text{lasso}}$  is;

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left( - \sum_{i=1}^n (y_i X_i^T \beta - \exp(X_i^T \beta) - \ln(y_i!)) + \lambda \sum_{i=1}^p |\beta_i| \right) \quad (4)$$

where  $\lambda$  is the Tuning Parameter of  $\hat{\beta}_{\text{lasso}}$ .

### 2.2.3 Poisson Ridge Regression

Kristofer and Ghazi (2011) developed a Poisson ridge regression. This method is popular and beneficial for the regression at the presence of multicollinearity problem by reducing the size of coefficients. The best parameters can be generated by using Log Likelihood function under L2 penalty (generalization) on  $\beta$ ,  $L_2 = \|\beta\|_2 = \sum_{i=1}^p \beta_i^2$ ; Then,  $\beta$  of the ridge regression can be estimated by the following equation;

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left( - \sum_{i=1}^n (y_i X_i^T \beta - \exp(X_i^T \beta) - \ln(y_i!)) + \lambda \sum_{i=1}^p \beta_i^2 \right) \quad (5)$$

where  $\lambda$  is the Tuning Parameter which controls the shrinkage size of  $\hat{\beta}_{\text{ridge}}$ .

### 2.2.4 Adaptive Poisson Lasso Regression

Park and Hastie (2007) developed the adaptive Poisson lasso regression as the extension of the Poisson regression model of Fan, Li (2001) that enhances estimation and variable selection performance by weighting method ( $w_i = 1/\beta_i$ ) for the reason that one variable can be more important than the others. We can find the coefficient by using Log Likelihood function under  $L_1$  penalty on  $\beta$ . The  $\hat{\beta}_{\text{Adaplasso}}$  can be calculated by the following equation;

$$\hat{\beta}_{\text{Adaplasso}} = \underset{\beta}{\operatorname{argmin}} \left( - \sum_{i=1}^n (y_i X_i^T \beta - \exp(X_i^T \beta) - \ln(y_i!)) + \lambda \sum_{j=1}^p |\beta_j| w_j \right)$$

where  $\lambda$  is the Tuning Parameter of  $\hat{\beta}_{\text{Adaplasso}}$ . However, it has some drawback regarding the Log Likelihood function since the best generated likelihood value might be overfitting when the number of parameters is large. To avoid using the overfitting parameters, we intend to replace the Log Likelihood approach by using Bayesian

### 3 Statistical Analysis

In this study, we use weekly data from over 70 economies in Asia (all countries in Asia and all provinces in China), spanning between March 16, 2020 and August 30, 2021 (5,865 observations). Our data set consists of daily COVID-19 cases from Our World in Data COVID-19 dataset and the policy responses by the Oxford Coronavirus Government Response Tracker (OxCGRT) such as Government Stringency Index, School and Workplace Closures, Cancellation of Public Events and Gatherings, Stay-at-Home Restrictions, Face Coverings, Public Information Campaigns, International and Domestic Travel, Testing and Contact Tracing, Vaccination Policy, Income Support, Debt Relief, etc. (see more detail in Hale et al. (2022), pp.30–42, the data we used includes both qualitative and quantitative variables, which are highly detailed). We also put variables like Economic Support Index, Median Age, Age 65 and older persons, GDP per capita, Cardiovascular Death Rate, and Diabetes Prevalence into our model to take into account the individual characteristic of each economy which may give us a better understanding of the policy efficacy in different contexts. We intend to elaborate on the policy measures that can help reduce COVID-19 infections within 2, 4, and 8 weeks after the implementation, using additive number of coefficients.

### 4 Empirical Results

In the first part of the estimation results, we would like to compare Poisson regression, Poisson ridge regression, Poisson lasso regression, and adaptive Poisson lasso regression using the BIC which is a criterion for model selection among a finite set of models where the lower value of BIC indicate a better fit model. The results regarding the BIC (Table 2) show that the Poisson lasso regression outperforms the Poisson ridge regression in our study by providing the lowest BIC value as the low BIC indicates a high performance of the forecasting model and vice versa.

**Table 1.** The comparison results.

	BIC
Poisson regression	624.53
Poisson lasso regression	<b>322.51</b>
Poisson ridge regression	2,511.44
Adaptive Poisson lasso regression	1,252.13

We chose to use Poisson lasso regression to show the approximation results in the next stage of the estimation. First and foremost, we chose total cases per million because we wanted to rescale the number of total cases, the dependent

variable, to be on the scale  $[0,1]$ , because countries with different sizes of populations may be affected differently by the policies. To avoid a misleading result, we intend to get rid of the effect of the size of the population which each country has different number of populations from the dependent variable, providing the result of per capital term. On the other hand, population size will be one of the explainable variables (Table 1).

For the understanding of the effect of policies, we will provide explanation by using the additive number of coefficients. The summation of coefficient values will give us the result of policies. If the summation of a policy is negative, it indicates that the policy helps cope with the total cases per million (we will state it shortly as total cases to be more compact). The autoregressive points out that the total case per million highly depends on its own past data; the closer, the greater the value. Since the positive rate of COVID-19 rises at the start of the pandemic, Asian governments usually implement policies to control the total number of cases per million. However, the effect of the positive rate, which indirectly causes the governments to handle the situation, will be diminishing, as we can see from the positive coefficients in lag 2 and lag 4.

An overview of high vaccination countries. Even the vaccination number is raising, however, it doesn't prevent COVID-19 spread. Moreover, it brings a higher total number of cases instead. One of the reasons behind this is that vaccines effectively protect people from death (Bloomberg, 2022), this would makes people relieve their anxiety, and the government can reduce the other countermeasures.

The evidence of total vaccinations is supported by the number of people fully vaccinated per hundred. After 4 weeks, the overall number of cases would have increased if everyone had been fully vaccinated. Even if the total number of vaccinations per hundred somehow doesn't help reduce the total cases, an increase in the number of vaccinated people per hundred helps reduce total cases after the policy has been implemented for 2–3 weeks. However, by one month, the impact had faded. Even while fully vaccinated people do not contribute to a reduction in cases, total boosters effectively reduce total cases within a month. People can usually acquire boosters after they have been fully vaccinated, and the effect of total booster numbers may be mitigated by total vaccinations and the number of people who have been fully vaccinated. The government should place a greater emphasis on the number of people vaccinated per hundred and total boosters when it comes to vaccination policy. Those policies, undoubtedly, aid in reducing the total number of cases.

The recommended closing, or all schools open with alterations resulting in significant differences compared to usual (School Closing1), is effective, and it helps reduce the number of new cases. Moreover, the required closing (School Closing2) takes effect within 3 weeks, but it provides better results than the recommended closing policy. Furthermore, the requirement to close all levels (School Closing 3, not shown in the table or graph) doesn't help reduce the total cases.



In workplace closing policies, the recommended closing policy or working from home (Workplace Closing1) is the most effective way compared to other workplace closing policies. The policy helped reduce the total number of cases in 3–4 weeks following the policy actions. However, the requirement to close all-but-essential workplaces (Workplace Closing3) reduced the total number of cases after it was applied for 4 weeks. However, the required closure of some sectors or categories of workers (Workplace Closing2) would exacerbate the overall situation.

The recommended canceling (Cancel Public Events1) helps deal with the increase in total cases after the policy has been in effect for a week, and it clearly outperforms the required canceling (Cancel Public Events2), which increased the total cases.

The restrictions on gatherings of 101–1000 people (Restrictions on Gatherings2) are the most effective compared to other policies. The restrictions on gatherings of 10 people or less policy (Restriction on Gatherings4) is effective just for the first week. However, the restriction on a gathering of 11–100 people (Restrictions on Gatherings3), usually employed in Australia and Saudi Arabia, is ineffective due to the positive number of the parameters. Moreover, the countries that implemented the restriction above 1,000 people (Restrictions on Gatherings1) benefited nothing from the policy.

The planned shutdown of public transit (Close Public Transport1: substantially lowering the volume, route, or method of transportation available) does not help to alleviate the problem; rather, it adds to it. However, after a month of implementation, the necessity to close public transit, which is more flexible, reduces the total number of occurrences.

The results of stay-at-home policies show that the more stringent the policy, the more effective it is in terms of time and effect comparing when people are completely confined at home (Stay at home requirements1: recommended not leaving the house) and prohibited from leaving the house except in exceptional circumstances (Stay at Home Requirements2). The recommended not leaving the house policy outperforms the policy with the fewest exceptions.

The recommend not to travel between regions/cities restriction (Restrictions on Internal Movement1) and the internal restriction policies (Restrictions on Internal Movement2) help control the total cases but the effect is small. While the internal movement restriction in place increases the total cases which were mentioned in the works of Quilty et al. (2020) and Bou-Karroum et al. (2021).

International travel control with quarantined arrivals from high-risk regions (International Travel Controls2) and bans on arrivals from some regions (International Travel Controls3) are effective ways to cope with the COVID cases. However, the countries that applied the tall border closure did obtain some positive effects on COVID-19 alleviation since the policy was applied for 7–8 weeks.

Next, the income support policies do not help reduce the total number of cases. Moreover, even if the government launches income support, the total number of cases still increases (in the case the government replaces less than 50% of the lost salary; Income Support1). The broad debt and contract relief

(Debt/Contract Relief of Households2) help reduce the total cases, obviously, and it takes effect within 2 weeks. While the narrow relief (Debt/Contract Relief of Households1: specific to one kind of contract) also helps cope with the situation, the result is outperformed by the broad relief and diminishing. Overall, the fiscal measures help decrease the total cases. However, the effect of the fiscal measures might be slower than 8 weeks due to the policy and spending transmission. Public officials help provide the information to people more effectively than the coordinated public information campaign (traditional and social media), which does not help decrease the cases.

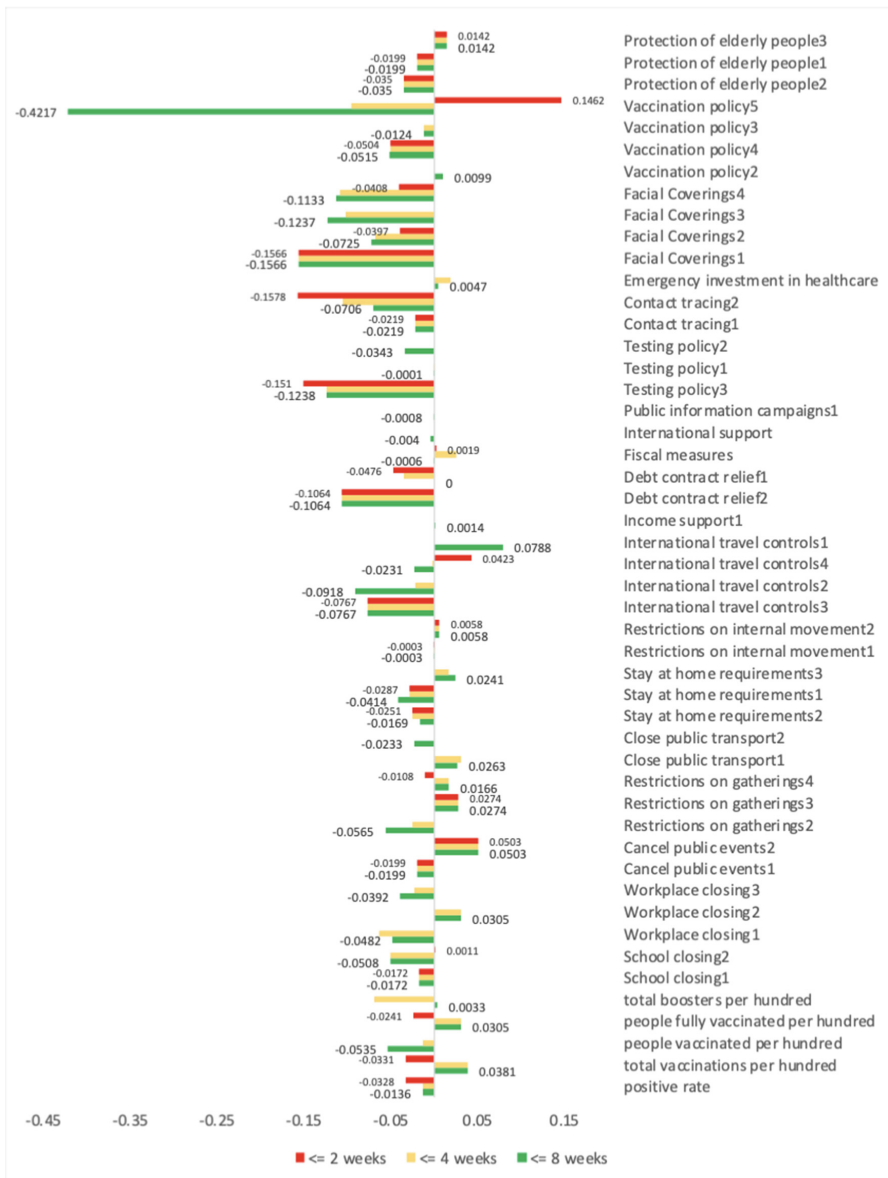
In terms of the testing policy, the open public testing policy (Testing Policy3) is the most effective, and the result is astonishing with a high coefficient value. It helps cope with the increase of total cases within a week. Moreover, the testing for anyone showing COVID-19 symptoms (Testing Policy2) decreases the total cases, but the effect is smaller and slower than the open public policy. However, the testing policy that is applied only to those who both have symptoms and meet specific criteria (Testing Policy1) shows a relatively small effect compared to the others. In the contact tracing policy, however, both the comprehensive and the limited contact tracing policies perform effectively in new case reduction. The emergency investment in healthcare doesn't provide a good effect within 8 weeks, the same as the investment in vaccines.

All facial covering policies are good solutions for COVID-19 pandemic alleviation. The results indicate that the lowest level of restriction, recommended facial coverings (Facial Covering1), is the most effective policy. While the requirement in all shared/public spaces or all situations (Facial Covering3) and required outside the home all the time (Facial Coverings4) provide slightly different results, and the requirement in some specified shared/public spaces outside the home (Facial Coverings2) gives the lowest effect but it is still a good policy. We notice that the stricter policies are usually employed in China, Egypt, Congo, Liberia, Mexico, and Spain (Our World in Data, 2022).

In the vaccination policy, the result shows that the universal availability policy, with no target group, is the best policy to cope with the new cases within 8 weeks (Vaccination Policy 5). Meanwhile Vaccination Policy4, availability for all three plus partial additional availability (select broad groups/ages), helps prevent new cases by providing vaccination for essential personnel, clinically fragile populations, geriatric groups, and partial additional availability. However, strategies that provide vaccination for some of the previously indicated target categories do not help minimize new cases; in fact, the availability of vaccine for two groups of workers (clinically susceptible group and the elderly) actually increases new cases. The specified and narrow restrictions are effective policies, while the extensive policies increase the number of new cases.

Lastly, the countries with higher median age, cardiovascular death rate, and diabetes prevalence tend to have more total cases per million. However, the countries with higher population over age of 65 and GDP per capita are developed countries which tend to have higher median age, tend to have less total cases per million. These findings show that our research is in line with those of earlier studies (Pian et al., 2021; Eibensteiner et al., 2021; Borio et al., 2022; Yang et

al., 2021) that described the causes, impacts, and preventative methods of the COVID-19 virus’s propagation efficiently (Fig. 1).



**Fig. 1.** The summation of parameter values: The percentage of reduction in total cases per million by policy implementations.

**Table 2.** Estimated parameters from the lasso model

Variable	Coef. value	Variable	Coef. value
(Intercept)	-0.5533	L8 International travel controls1	0.0747
L1 total cases per million	0.8156	L8 International travel controls4	-0.0035
L2 total cases per million	0.0733	L6 Income support1	0.0014
L3 total cases per million	0.0264	L1 Debt contract relief1	-0.0476
L5 total cases per million	0.0218	L2. Debt contract relief2	-0.1064
L6 total cases per million	0.0175	L4 Debt contract relief1	0.0118
L7 total cases per million	0.0047	L5 Debt contract relief1	0.0074
L8 total cases per million	0.0091	L6 Debt contract relief1	0.0283
L1 positive rate	-0.0665	L8 Debt contract relief1	0.0001
L2 positive rate	0.0337	L1 Fiscal measures	0.0019
L4 positive rate	0.0192	L3 Fiscal measures	0.0106
L1 total vaccinations per hundred	0.0382	L4 Fiscal measures	0.0124
L2 total vaccinations per hundred	-0.0001	L5 Fiscal measures	-0.0195
L1 people vaccinated per hundred	0.0001	L7 Fiscal measures	-0.0057
L2 people vaccinated per hundred	-0.0332	L8 Fiscal measures	-0.0003
L3 people vaccinated per hundred	-0.0003	L6 International support	-0.0040
L4 people vaccinated per hundred	0.0200	L8 Public information campaigns1	-0.0008
L8 people vaccinated per hundred	-0.0401	L1 Testing policy3	-0.1510
L1 people fully vaccinated per hundred	0.0041	L3 Testing policy3	0.0272
L2 people fully vaccinated per hundred	-0.0282	L4 Testing policy1	-0.0001
L4 people fully vaccinated per hundred	0.0546	L5 Testing policy1	0.0000
L3 total boosters per hundred	-0.0695	L7 Testing policy2	-0.0343
L5 total boosters per hundred	0.0773	L1 Contact tracing1	0.0492
L6 total boosters per hundred	-0.0364	L1 Contact tracing2	-0.1578
L8 total boosters per hundred	0.0319	L2 Contact tracing1	-0.0711
L1 School closing1	-0.0172	L4 Contact tracing2	0.0517
L1 School closing2	0.0011	L5 Contact tracing2	0.0355
L3 School closing2	-0.0519	L3 Emergency investment in healthcare	0.0184
L3 Workplace closing1	-0.0547	L5 Emergency investment in healthcare	-0.0102
L4 Workplace closing1	-0.0093	L6 Emergency investment in healthcare	-0.0035
L4 Workplace closing2	0.0305	L1 Facial Coverings2	-0.0288
L4 Workplace closing3	-0.0234	L2 Facial Coverings1	-0.1566
L5 Workplace closing3	-0.0108	L2 Facial Coverings2	-0.0397
L7 Workplace closing1	0.0158	L2 Facial Coverings4	-0.0408
L8 Workplace closing3	-0.0050	L3 Facial Coverings3	-0.1019
L1 Cancel public events1	-0.0199	L4 Facial Coverings4	-0.0682
L1 Cancel public events2	0.0503	L5 Facial Coverings3	-0.0218
L1 Restrictions on gatherings3	0.0274	L8 Facial Coverings2	-0.0040
L1 Restrictions on gatherings4	-0.0108	L8 Facial Coverings4	-0.0043
L3 Restrictions on gatherings2	-0.0142	L1 Vaccination policy4	0.1771
L4 Restrictions on gatherings2	-0.0112	L1 Vaccination policy5	0.7908
L5 Restrictions on gatherings2	-0.0260	L2 Vaccination policy4	-0.2275
L8 Restrictions on gatherings2	-0.0325	L2 Vaccination policy5	-0.6446
L3 Close public transport1	0.0304	L3 Vaccination policy5	-0.6573
L5 Close public transport2	-0.0476	L4 Vaccination policy3	-0.0124
L7 Close public transport1	-0.0016	L4 Vaccination policy5	0.4154
L7 Close public transport2	0.0243	L5 Vaccination policy2	0.0072
L8 Close public transport1	-0.0025	L5 Vaccination policy4	-0.0012
L1 Stay at home requirements2	-0.0251	L6 Vaccination policy2	0.0444
L2 Stay at home requirements1	-0.0287	L6 Vaccination policy5	-0.3260
L3 Stay at home requirements3	0.0168	L8 Vaccination policy2	-0.0417
L5 Stay at home requirements1	-0.0127	L8 Vaccination policy4	0.0001
L5 Stay at home requirements2	0.0082	L1 Protection of elderly people1	0.0023
L8 Stay at home requirements3	0.0073	L1 Protection of elderly people2	-0.0350
L1 Restrictions on internal movement1	-0.0003	L2 Protection of elderly people1	-0.0222
L2 Restrictions on internal movement2	0.0058	L2 Protection of elderly people3	0.0142
L2 International travel controls3	-0.0767	EconomicSupportIndex	0.0010
L2 International travel controls4	0.0423	Median age	0.2778
L3 International travel controls2	-0.0218	Aged 65 older	-0.0133
L4 International travel controls4	-0.0445	GDP per capita	-0.0058
L5 International travel controls2	-0.0567	Cardiovascular death rate	0.0015
L6 International travel controls1	0.0041	Diabetes prevalence	0.2212
L6 International travel controls2	-0.0133		
L7 International travel controls4	-0.0174		

Note 1: L indicates lag number, e.g., L1 is a lag 1, L2 is a lag 2

Note 2: This table shows only the selected variables.

## 5 Conclusion

The spread of the COVID-19 pandemic in 2020 has exerted a large impact on various economic sectors. In this paper, we analyze the policies that were implemented during the epidemic, such as school and workplace closures, cancellation of public events and gatherings, stay-at-home restrictions, face coverings, public information campaigns, international and domestic travel, testing and contact tracing, vaccination policy, etc. we use Poisson lasso model as it is very easy to apply. This model includes a relationship the best variable selection and the connections between the lasso coefficient estimates.

The results showed that the lasso regression chose 58 variables, each with a different direction and magnitude, that affected the spread of COVID. Certain policies or variables can reduce the total number of infections one week after they are adopted. The best method to cope with the outbreak is to implement the universal vaccination policy (vaccination policy 5). However, several strategies, such as the vaccination policy (vaccination policy 2), testing policy (testing policy 2: testing anyone showing COVID-19 symptoms), public information campaigns, international support, and closing public transportation (require closing or prohibit most citizens from using it), are relatively slow, taking 1 to 2 months to reduce the number of cases. Meanwhile, policies such as cancelling public events (require cancelling), international travel controls (screening), workplace closure (require closing for some sectors or categories of workers), restrictions on gatherings (restriction on gatherings of 10 people or less), closing public transportation (recommend closing or significantly reduce volume/route/means of transport available), and stay-at-home requirements (require not leaving house with minimal exceptions) have all contributed to an increase in the number of infections.

Therefore, dealing with the epidemic requires choosing a policy that is appropriate for the current situation and taking into account the consequences that will determine whether they will be able to reduce the number of the infected or not. However, this study is only examining policies that emerged during the epidemic based on the perspective of the lasso regression. It does not take into account other methods that may show better results and perspectives, for example, the vector autoregression (VAR) model. Poisson lasso regression is used in this study to describe the influence of various policy measures on the number of infections using our data set; but the method is also believed to be able to improve the estimation results using other data sets.

**Acknowledgements.** This research work was partially supported by Chiang Mai University

## References

Faisal, M., Nirmala, M.P.: COVID-19 and economic policy options: what should the government do?. *Jurnal Inovasi Ekonomi* 5(02) (2020)

- Sharif, A., Aloui, C., Yarovaya, L.: COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: fresh evidence from the wavelet-based approach. *Int. Rev. Financ. Anal.* **70**, 101496 (2020)
- Baker, S.R., Bloom, N., Davis, S.J., Kost, K. J., Sammon, M.C., Viratyosin, T.: The unprecedented stock market impact of COVID-19 (No. w26945) (2020). National Bureau of Economic Research
- Baker, S.R., Bloom, N., Davis, S.J., Terry, S.J.: COVID-induced economic uncertainty (No. w26983). National Bureau of Economic Research (2020)
- Conlon, T., McGee, R.: Safe haven or risky hazard? Bitcoin during the COVID-19 bear market. *Financ. Res. Lett.* **35**, 101607 (2020)
- Corbet, S., Larkin, C., Lucey, B.: The contagion effects of the COVID-19 pandemic: evidence from gold and cryptocurrencies. *Financ. Res. Lett.* **35**, 101554 (2020)
- Caporale, G.M., You, K., Chen, L.: Global and regional stock market integration in Asia: a panel convergence approach. *Int. Rev. Financ. Anal.* **65**, 101381 (2019)
- Wu, L., Meng, Q., Xu, K.: 'Slow-burn' spillover and 'fast and furious' contagion: a study of international stock markets. *Quant. Finan.* **15**(6), 933–958 (2015)
- Dimitriou, D., Kenourgios, D., Simos, T.: Global financial crisis and emerging stock market contagion: a multivariate FIAPARCH-DCC approach. *Int. Rev. Financ. Anal.* **30**, 46–56 (2013)
- Fetzer, T., Hensel, L., Hermle, J., Roth, C.: Coronavirus perceptions and economic anxiety. *Rev. Econom. Stat.* 1–36 (2020)
- Ma, C., Rogers, J.H., Zhou, S.: Global economic and financial effects of 21st century pandemics and epidemics. In: SSRN (2020)
- Yarovaya, L., Matkovskyy, R., Jalan, A.: The effects of a 'Black Swan' event (COVID-19) on herding behavior in cryptocurrency markets: evidence from cryptocurrency USD, EUR, JPY and KRW Markets. *EUR, JPY and KRW Markets* 27 April 2020 (2020)
- Alon, T.M., Kim, M., Lagakos, D., VanVuren, M.: How should policy responses to the COVID-19 pandemic differ in the developing world? (No. w27273). National Bureau of Economic Research (2020)
- Hsiang, S., et al.: The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**(7820), 262–267 (2020)
- Whitelaw, S., Mamas, M.A., Topol, E., Van Spall, H.G.: Applications of digital technology in COVID-19 pandemic planning and response. *Lancet Digit. Health* **2**, e435–e440 (2020)
- World Bank. The COVID-19 Pandemic: Shocks to Education and Policy Responses (2020)
- Hevia, C., Neumeyer, A.: A conceptual framework for analyzing the economic impact of COVID-19 and its policy implications. UNDP LAC COVID-19 Pol. Doc. Ser. **1**, 29 (2020)
- McBryde, E.S., et al.: Role of modelling in COVID-19 policy development. *Paediatric Respir. Rev.* **35**, 57–60 (2020)
- Dev, S.M., Sengupta, R.: COVID-19: Impact on the Indian Economy. Indira Gandhi Institute of Development Research, Mumbai (2020)
- Bou-Karroum, L., et al.: Public health effects of travel-related policies on the COVID-19 pandemic: a mixed-methods systematic review. *J. Infect.* **83**(4), 413–423 (2021)
- Quilty, B.J., et al.: The effect of travel restrictions on the geographical spread of COVID-19 between large cities in China: a modelling study. *BMC Med.* **18**(1), 1–10 (2020)

- Hale, T., et al.: Variation in government response to COVID-19. BSG Working Paper Series (2022). <https://www.bsg.ox.ac.uk/sites/default/files/2022-04/BSG-WP-2020-032-v13.pdf>
- Pian, W., Chi, J., Ma, F.: The causes, impacts and countermeasures of COVID-19 “Infodemic”: a systematic review using narrative synthesis. *Inf. Proc. Manage.* **58**(6), 102713 (2021)
- Eibensteiner, F., et al.: Countermeasures against COVID-19: how to navigate medical practice through a nascent, evolving evidence base—a European multicentre mixed methods study. *BMJ Open* **11**(2), e043015 (2021)
- Borio, L.L., Bright, R.A., Emanuel, E.J.: A national strategy for COVID-19 medical countermeasures: vaccines and therapeutics. *JAMA* **327**(3), 215–216 (2022)
- Yang, F., Zhang, J., Zheng, W., Huang, M., Zhang, L., Zhang, B.: New problems and countermeasures in the prevention and control of COVID-19. *Emerg. Crit. Care Med.* **1**(1), 12–13 (2021)



# Correlation Between Foreign Ownership and Liquidity Risk

Nguyen Ngoc Thach<sup>1</sup> , Bui Dan Thanh<sup>2</sup>, and Le Thi Lan<sup>3</sup> 

<sup>1</sup> Institute for Research Science and Banking Technology, Banking University HCMC,  
36 Ton That Dam, District 1, Ho Chi Minh City, Vietnam

thachnn@buh.edu.vn

<sup>2</sup> Banking Technology, Banking University HCMC, 36 Ton That Dam, District 1,  
Ho Chi Minh City, Vietnam

thanhbd@buh.edu.vn

<sup>3</sup> Joint Stock Commercial Bank for Investment and Development of Vietnam,  
137C Nguyen Chi Thanh St., Ward 9, District 5, Ho Chi Minh City, Vietnam

lanbt6988@gmail.com

**Abstract.** The article studies the influence of foreign ownership ratio on liquidity risk of Vietnamese commercial banks in the period 2009–2020. The article uses regression methods based on Bayesian approach with sample data of 30 Vietnamese commercial banks. The research results show that the higher the foreign ownership ratio, the lower the liquidity risk of commercial banks, as expected for the study. Besides, the variables of credit risk, equity ratio, loan-to-deposit ratio and economic growth have significant impact on liquidity risk.

**Keywords:** Foreign ownership · liquidity risk · commercial banks · Bayesian regression · Vietnam

## 1 Introduction

In the integrated and developed economy nowadays, commercial banks are focusing on finding strategic partners to develop their business and minimize risks specific to the banking industry. In addition, signed free trade agreements make access to foreign capital increasingly easier.

The amount of foreign ownership in Vietnamese banks accounts for a large market share (more than 60% of Vietnamese commercial banks have capital from foreign investors). However, in many banks, the foreign ownership ratio is quite modest. In addition, according to Decree 01/2014/ND-CP limiting the foreign ownership ratio not to exceed 30% of the capital of commercial banks in Viet Nam, the opening of the door to welcome foreign investors is very limited.

However, in terms of academics, there are not many practical studies on this issue, or only revolve around making profits without strongly focusing on risk management. Therefore, the question is whether foreign ownership actually performs well in liquidity risk management at commercial banks of Vietnam. To answer the above question,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 574–587, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_41](https://doi.org/10.1007/978-3-031-35763-3_41)



we chose the topic: “The impact of foreign ownership on liquidity risk of Vietnamese commercial banks”.

## 2 Literature Review

### 2.1 Theory of Foreign Ownership

Foreign ownership, also known in specialized terms, is the percentage (%) of shares held by all foreign investors. According to Decree 01/2014/ND-CP: “The total share ownership of foreign investors must not exceed 30% of the charter capital of a Vietnamese commercial bank”.

According to previous studies, the foreign ownership ratio is determined by the following formula:

$$\text{FOREIGN} = \frac{\text{Shares of foreign shareholders}}{\text{Total number of shares issued}}$$

### 2.2 Liquidity Theory of Commercial Banks

According to Duttweiler (6), liquidity is the ease with which a particular asset is converted into cash and the market accepts that transaction. Bank liquidity includes two types: natural liquidity and artificial liquidity. Liquidity risk occurs when the bank is short of short-term assets with high liquidity such as cash, gold, silver, precious stones, deposits at the State bank or other credit institutions, etc.... to meet the needs of depositors and borrowers.

### 2.3 Theory of Liquidity Risk

According to Circular 08/2017/TT-NHNN, liquidity risk is defined as follows: “Liquidity risk is the risk that credit institutions, foreign bank branches are unable to fulfill their obligations to repay the debt when it is due; or a credit institution or foreign bank branch that is capable of performing a debt repayment obligation when it is due, but has to pay high costs to fulfill that obligation.”

According to Duttweiler (6), liquidity risk is defined as the risk arising when a commercial bank is no longer able to pay at a certain time, or has to raise capital from a third party at high cost to meet the demand of instant payment.

According to Athanasoglou et al. (2); Demirgüç-Kunt and Huizinga (4); Tran Hoang Ngan and Pham Quoc Viet (2016) liquidity risk is measured using the formula:

$$L_3 = \frac{\text{Loans}}{\text{Total Assets}}$$

The higher the ratio, the higher the bank’s liquidity risk, which means that the higher the bank’s lending ratio, the higher its liquidity risk.

## 2.4 Comprehensive Researches

Nguyen et al. (11) investigates how foreign ownership and management affect listed companies' financial performance in the Vietnamese stock market. 427 listed companies from all industries were included in the data throughout a five-year period, from 2014 to 2018. ROA, and ROE are used to gauge a company's financial success. The study tested each model using the Pool OLS least squares approach while also considering fixed effects (FEM), random effects (REM). The FEM model is the most practical one. The findings indicate that the size of the company and the percentage of foreign ownership have a favorable effect on financial success. Financial performance is negatively impacted by foreign management, age of the companies, liquidity, and financial leverage.

Kusi et al. (12) uses information on 26 banks that was gathered between 2006 and 2016 from the Bank of Ghana. To arrive at the results, three panel estimation strategies: two-step GMM, Hausman-Taylor and Fixed effect models were used. Regression models are used in the study, which reveals that foreign and privately owned banks are less likely to produce more liquidity than their domestic and state-owned bank counterparts, suggesting that domestic and state-owned banks produce more liquidity. These findings suggest that although there is a lot of room for more liquidity to be created, policymakers may speed up the process by using state- and locally-owned banks while also designing policies that encourage foreign and privately owned banks to increase their liquidity creation, which is beneficial for economic growth.

Le (13) investigates the impact of foreign ownership on bank risk in Vietnam between 2006 and 2015. The findings suggest that the State Bank of Vietnam should further relax its limitations on foreign investments in the banking system since foreign ownership can reduce bank risk. The results also show a relationship between bank risk and technological efficiency, suggesting the existence of the skimping-cost hypothesis. The same finding holds true for big banks, institutions with more liquid assets, and institutions with faster loan growth. According to the author's findings, state-owned banks with higher levels of foreign ownership are probably more stable. The same holds true for listed banks that have a bigger percentage of foreign ownership.

Al-Harbi (14) investigate the determinants of Islam banks (IBs) liquidity. On an imbalanced panel data set of all IBs operating in the nations of the Organization of Islamic Cooperation from 1989 to 2008, the author applies a generalized least square fixed effect model. All of the factors have statistically significant correlations with IBs' liquidity, according to the estimation results, but these relationships have distinct signs. On the one hand, IBs' liquidity was adversely impacted by foreign ownership, credit risk, profitability, inflation rate, monetary policy, and deposit insurance. On the other hand, there is a strong correlation between the liquidity of IBs and the capital ratio, size gross domestic product growth and concentration.

Nacerayeddou et al. (2020) examine the connection between bank ownership structure and bank liquidity creation for the years 2004–2018 using a new, hand-collected database on ownership structure for a sample of commercial banks from 17 western European nations. The concentration of bank ownership and the identity of the principal owner are the authors' main concerns. The effects are twofold: first, ownership concentration significantly and favorably affects the generation of liquidity. Analyze the

effect of the owner's nature on the creation of liquidity next. When another bank or the government owns more than 50% of a bank, 65% of a non-financial company, 75% of a family, or 85% of a financial organization, banks tend to create more liquidity, according to the authors.

### 3 Model and Method

From the review of previous studies, the research team proposes the following research hypotheses:

**FOREIGN<sub>i,t</sub> – Foreign Ownership Ratio:** According to Terrell (8), foreign owned banks can indirectly increase efficiency by stimulating competition in the domestic financial market. In addition, foreign-owned banks have improved their supervisory and regulatory frameworks, lending quality and risk management. Therefore, this study expects that the higher the foreign ownership ratio, the lower the bank's liquidity risk.

**Hypothesis 1: There is a negative relationship between foreign ownership ratio and liquidity risk ( $H_1$ ).**

**CR<sub>i,t</sub> – Credit Risk:** Banks in Vietnam are focusing mainly on lending activities and have a high bad debt ratio, the higher the level of bad debt, the more provisions the bank makes, that is, as the provisioning increases, the bank's profit accordingly decreases. In order to ensure profitability, banks tend to lend more and cut down on highly liquid assets. This means that when credit risk increases, the bank's liquidity risk increases.

**Hypothesis 2: There is a positive relationship between liquidity risk and credit risk ( $H_2$ ).**

**SIZE<sub>i,t</sub> – Bank Size:** According to most authors, bank size always affects liquidity risk in two directions, either positive or negative. If SIZE has a positive correlation with liquidity risk, it shows that if the scale is expanded, the operating and management costs will increase, human resources are not enough to control the risk. If SIZE has a negative correlation with liquidity risk, it means that the more the bank expands, the more likely the bank will be able to attract capital sources, as well as lend more and bring in more profits for the bank. Due to expansion, it is easier to attract external funds to meet short-term liquidity needs in a timely manner, meaning liquidity risk is reduced. **Hypothesis 3: There is a negative relationship between liquidity risk and bank size ( $H_3$ ).**

**EQUITY<sub>i,t</sub> – Equity Ratio:** According to the basic hypothesis of return and risk is "High risk high return", that is, taking risks will receive a larger return, which means if this ratio is low, the bank's profits increase by taking on a moderate level of risk. According to Circular 41/2016/TT-NHNN and Circular 22/2019/TT-NHNN regulating capital adequacy ratio. Accordingly, in order to meet the CAR ratio, banks are racing to increase their own capital. When a bank has a large capitalization, the capital adequacy ratio and liquidity ratio will also increase, which means the bank's liquidity risk will decrease. **Hypothesis 4: There is a negative relationship between liquidity risk and equity ratio ( $H_4$ ).**

**LDR<sub>i,t</sub> – Lending/Depositing Ratio:** According to Golin (7), a higher ratio means more loans than mobilized capital. Therefore, when facing liquidity risk, it will be

difficult for banks to mobilize cheap capital if they lend too much, reducing the bank's liquidity, which means increased liquidity risk. Also according to the author, when this ratio is low, banks can easily mobilize from various sources such as interbank market, issue of valuable papers, etc. with cheap capital, making the liquidity of banks increase.

**Hypothesis 5: There is a positive relationship between liquidity risk and loan/deposit ratio ( $H_5$ ).**

**ROA<sub>i,t</sub> – Profit/Total Assets:** Profit after tax after one year of a bank is used for two main purposes: retained earnings for reinvestment and/or distribution of profits to shareholders. When profits are retained, they are also reinvested in a bank account. When the profit/total assets ratio is high, it means that the bank's liquidity is high, which means that the liquidity risk is low (Aspachs, 1). **Hypothesis 6: There is a negative relationship between liquidity risk and return/total assets ( $H_6$ ).**

**DR<sub>t</sub> – Average Real Deposit Rate:** When the bank's deposit rate decreases, the deposit flow will move to a place with higher interest rates. At that time, the domino effect will take place, causing customers to suddenly withdraw their deposits but other loans and receivables have not been due to be settled, causing the bank to temporarily lose liquidity. When deposit interest rates are high, banks will limit their holding of highly liquid and low-profit assets because those assets are not profitable enough for the bank to ignore. This increases the bank's liquidity risk. **Hypothesis 7: There is a negative relationship between liquidity risk and average real deposit rate ( $H_7$ ).**

**IR<sub>t</sub> – Real Interbank Interest Rate:** According to Dinger (5), real interbank interest rate is an index to measure liquidity costs in the banking system. The real interbank rate is determined as the net value between the 1-month interbank rate and annual inflation. When banks need liquidity to pay their due debts, banks can mobilize capital from external sources with high interest rates but can also borrow through the interbank market with cheap capital. Therefore, the interbank interest rate reflects the liquidity status of the banking system and is continuously updated by the central bank. **Hypothesis 8: There is a positive relationship between liquidity risk and real interbank interest rate ( $H_8$ ).**

**SMR<sub>t</sub> – market Interest Rate Volatility Index:** According to Dinger (5), the market interest rate volatility index is measured by the standard deviation of the 1-month term interbank market interest rate, this index is given on the liquidity shortage of the whole banking system. Thereby, investors and policy makers can observe the situation of the currency market. According to Von Hagen and Ho (10); Dinger (5) research shows that the market interest rate and the liquidity situation of the banking system have an inverse relationship, that is, when the market interest rate decreases, the liquidity of the banking system is good and risky liquidity is minimized. **Hypothesis 9: There is a positive relationship between liquidity risk and market interest rate volatility index ( $H_9$ ).**

**GDP<sub>t</sub> – Economic Growth:** In good and stable economic conditions, people will have excess capital and save more, the liquidity of banks will be stable. But on the contrary, when the economy is exhausted, loans with bad debts are difficult to recover, affecting debt recovery. When the payables are due, the bank's liquidity is not enough to meet the customer's withdrawal demand, and the liquidity risk increases. **Hypothesis 10: There is a negative relationship between liquidity risk and economic growth ( $H_{10}$ ).**

**NIM<sub>t</sub> – the Difference Between Lending and Deposit Rates in the Industry:** According to Aspachs et al. (1); Vodova (9); Bonfim and Kim (3) introduced the difference between lending interest rates and deposit rates of the whole industry as a new point for the research topic. When this difference is high, the amount of money mobilized is less and the loan disbursement is also less (because the deposit interest rate is quite low while the lending interest rate is quite high). But if this difference is low, it will affect the bank's profit. According to Vodova (9), NIM does not affect the liquidity of banks. But Bonfim and Kim (3) found that NIM and liquidity risk were inverse, while Aspachs et al (1) showed the same results. When NIM increases, it means that banks earn more money, which means that the bank's ROA also increases (Table 1). ***Hypothesis 11: The difference between lending interest rates and deposit interest rates in the whole industry exists in the opposite direction. Liquidity risk (H<sub>11</sub>).***

Thus, the research model has the form:

$$LR = \beta_1 \text{FOREIGN} + \beta_2 \text{CR} + \beta_3 \text{SIZE} + \beta_4 \text{EQITY} + \beta_5 \text{LDR} + \beta_6 \text{ROA} + \beta_7 \text{DR} + \beta_8 \text{IR} + \beta_9 \text{SMR} + \beta_{10} \text{GDP} + \beta_{11} \text{NIM} + \varepsilon$$

The study is based on unbalanced panel data. The data is compiled from Financial Statements, Annual Reports of 30 commercial banks for the period from 2009 to 2020.

To conduct a Bayesian analysis, a priori information is required for the research model, but since most of the prior research was performed using a frequency approach, a priori information is not available. However, with the research data of 30 banks in the period 2009–2020, the number of observations is very large, so the priori information does not have a great influence on the posterior distribution. In this case, Block et al. (2011) proposed a standard Gaussian distribution with different a priori information (simulation of a priori information) and carried out Bayesian factor analysis to choose a simulation with the best previous information.

The simulations in Table 2 show decreasing levels of a priori information with Simulation 1 having the strongest a priori information and Simulation 5 having the weakest a priori information.

In the next step, the author carried out Bayesian regression for the above simulations, then performed Bayesian factor analysis and Bayestest model. These are the techniques proposed by StataCorp LLC (2019) to select the simulation with the best a priori information. Basically, the Bayesian factor will provide a tool to compare the probability of a particular hypothesis (a priori information) to the probability of another hypothesis. It can be understood as a measure of the strength of evidence in favor of a theory among competing (information a priori) theories. Accordingly, Bayesian analysis will provide average Log BF (Bayes Factor - Bayes factor), Log ML (Marginal Likelihood - marginal likelihood) and average DIC (Deviance Information Criterion - information bias); The posterior Bayesian test will help compare the posterior probability of the simulations with different a priori information, accordingly, based on the research data combined with the proposed a priori information, we will choose The simulation has the greatest posterior probability P(Mly).

In summary, in this study, the research team will build 5 simulations with 5 different a priori information, and Bayesian factor analysis and posterior Bayes test will help to choose a simulation with suitable a priori information. The simulation selected will be

**Table 1.** The data used in the research model

Description	Variable	Formula	Expectation
Dependent variable			
Liquidity risk	LR	$\frac{\text{Loans}}{\text{Total Assets}}$	
Independent variable			
Foreign ownership ratio	FOREIGN	$\text{FOREIGN} = \frac{\text{Shares of foreign shareholders}}{\text{Total number of shares issued}}$	-
Credit risk	CR	$\frac{\text{Provision for credit risks}}{\text{Total Assets}}$	+
Bank size	SIZE	Log (Total Assets)	-
Equity ratio	EQUITY	$\frac{\text{Equity}}{\text{Total Assets}}$	-
Lending/depositing ratio	LDR	$\frac{\text{Lending}}{\text{Depositing}}$	+
Profit to total assets	ROA	$\frac{\text{EAT}}{\text{Total Assets}}$	-
Average real deposit rate	DR	12-month term deposit interest rate – Annual inflation	-
Real Interbank Interest Rate	IR	1-month term interbank interest rate – Annual inflation	+
Market interest rate volatility index	SMR	$\sqrt{\frac{1}{n} \sum_{i=1}^n (\text{IR}_i - \overline{\text{IR}})^2}$	+
Economic growth	GDP	Log (GDP)	-
The difference between lending and deposit rates in the industry	NIM	Loan interest rate – Deposit interest rate	-

Note: + is the positive effect, - is the opposite effect

Source: Compiled by the author

the one with the largest Log BF, Log ML average, minimum DIC mean and the largest P(Mly).

## 4 Research Results and Discussion

Table 3 shows that simulation 1 meets the criteria to be the most suitable priori information simulation. Moreover, the results of post-test also show that simulation 1 has superiority over other simulations, so simulation 1 with a priori information  $N(0, 1)$  will be selected.

Bayes analysis is simulated through the Markov chain Monte Carlo (MCMC), therefore, to ensure the stability of the Bayesian regression, the MCMC series must converge, which means that the MCMC series must ensure stationarity. StataCorp LLC (2019) proposes that the MCMC series convergence test can be conducted through the convergence diagnostic graph.

**Table 2.** Simulation of a priori information

Rational function	$LR \sim N(\mu, \sigma)$
A priori distribution	
Simulation 1	$\alpha \sim N(0, 1)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 2	$\alpha \sim N(0, 10)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 3	$\alpha \sim N(0, 100)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 4	$\alpha \sim N(0, 1000)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$
Simulation 5	$\alpha \sim N(0, 10000)$ $\sigma^2 \sim \text{Invgamma}(0.01, 0.01)$

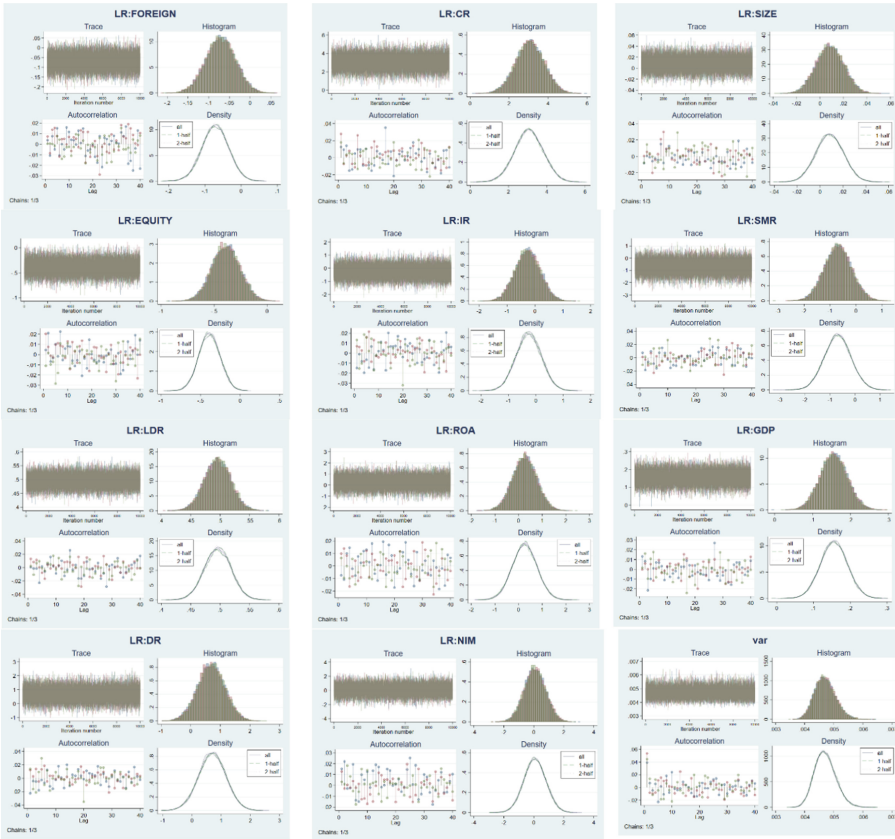
Source: Compiled by the author

**Table 3.** Bayes Factor analysis results

	Chains	Avg DIC	Avg log (ML)	Log (BF)	P (Mly)
SALEG1	3	-884.4302	409.2040		0.9299
SALEG2	3	-891.2087	406.6183	-2.5857	0.0701
SALEG3	3	-890.8738	394.7687	-14.4353	0
SALEG4	3	-890.7863	381.2600	-27.9440	0
SALEG5	3	-890.7615	367.4591	-41.7450	0

Source: Calculations of the author

According to StataCorp LLC (2019), the MCMC series convergence diagnostic graph includes trace plot, histogram, autocorrelation, and density plot. The trace plot helps to track the historical display of a parameter value over the iterations of the series, Fig. 1 shows the trace plot fluctuates around the mean value, so the MCMC series is stationary, that is, reaching convergence conditions. Besides, the autocorrelation chart in the graphs only fluctuates around the level below 0.02, according to StataCorp LLC (2019) the autocorrelation chart fluctuates around the level below 0.02, showing the agreement with the density the distribution and reflect all delays that are within the effective limit. According to StataCorp LLC (2019), the posterior distribution plot and density estimate show that the simulation of the shape of the normal distribution of the parameters, the histogram shape is uniform, it can be concluded that Bayes regression ensure stability. Thus, the results from Fig. 1 show that the MCMC series meets the convergence condition.



**Fig. 1.** Convergence diagnostic graph. Source: Calculations of the author

In addition to graphical convergence diagnostics, StataCorp LLC (2019) also recommends testing through Mean Acceptance Rate; Average minimum efficiency; and Gelman-Rubin  $R_c$  max. Table 4 shows that the model's acceptance rate reaches 1, the model's minimum efficiency is 0.91, far exceeding the allowable level of 0.01; In addition, the maximum  $R_c$  value of the coefficients is 1, Gelman and Rubin (1992) argue that the diagnostic value  $R_c$  of any coefficient of the model greater than 1.2 will be considered non-convergent. Thus, the values in Table 4 show that the MCMC series of the model satisfy the convergence requirements.

Regression results in the Table 4 have identified the variables FOREIGN and EQUITY have negative impact on liquidity risk (LR) while the variable CR, LDR, GDP increase liquidity risk. Besides determining the sign of the regression coefficients, unlike the frequency method, the Bayes approach also allows us to calculate the probability of the occurrence of these effects (Table 5).

The results show the probability that when the foreign ownership ratio (FOREIGN) is higher, the liquidity risk (LR) of Vietnamese joint stock commercial banks tends to decrease, that is, the foreign ownership ratio tends to increase with a probability of



**Table 4.** Regression results

	Mean	Std. Dev	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
FOREIGN	-0.0712	0.0364	0.0002	-0.0712	-0.1426	-0.0004
CR	3.0656	0.7403	0.0044	3.0649	1.6029	4.5038
SIZE	0.0085	0.0119	0.0001	0.0084	-0.0147	0.0318
EQUITY	-0.3879	0.1340	0.0008	-0.3885	-0.6485	-0.1260
IR	-0.2617	0.4573	0.0027	-0.2617	-1.1499	0.6380
SMR	-0.6947	0.5305	0.0031	-0.6949	-1.7283	0.3508
LDR	0.4958	0.0224	0.0001	0.4959	0.4517	0.5395
ROA	0.2569	0.5193	0.0030	0.2585	-0.7609	1.2750
GDP	0.1556	0.0364	0.0002	0.1555	0.0846	0.2270
DR	0.6908	0.4647	0.0027	0.6922	-0.2340	1.6003
NIM	0.0702	0.7317	0.0043	0.0649	-1.3604	1.5042
_cons	-1.6052	0.3997	0.0023	-1.6043	-2.3940	-0.8202
var	0.0047	0.0004	0.0000	0.0047	0.0040	0.0055
Avg acceptance rate	1					
Avg efficiency min	0.9085					
Max Gelman-Rubin Rc	1					

Source: Calculations of the author

more than 97%. This result is consistent with the study of Hammami and Boubaker (2015); Laeven (1999); Demirgüç-Kunt and Huizinga (4). According to Terrell (8), by promoting competition in the domestic financial sector, foreign-invested commercial banks can indirectly promote efficiency. The supervisory and regulatory framework, lending standards and risk management of commercial banks have all been strengthened by commercial banks with foreign capital. Therefore, the smaller the liquidity risk of commercial banks, the greater their foreign ownership ratio. Research results show that foreign partners have a very good impact on liquidity management to reduce risks.

In addition, equity ratio (EQUITY) also has a negative effect on liquidity risk (LR), equity ratio tends to increase with a probability of more than 99%. This result is consistent with research expectations and in agreement with the studies of Bunda and Desquilbet (2008); Lucchetta (2007); Vodova (9); Vu Thi Hong (2015). Commercial banks tend to reduce risk when equity ratio is higher because the source of money for business operations is equity, which will affect investment and lending regulations. Commercial banks must ensure capital adequacy ratio as prescribed in Circular 41/2016/TT-NHNN and Circular 22/2019/TT-NHNN regulating capital adequacy ratio. Commercial banks are preparing to increase their own capital to meet the CAR requirement. The liquidity risk of a commercial bank will be reduced if it is highly capitalized as it will have a higher capital adequacy ratio and liquidity ratio.

**Table 5.** Probabilistic test

	Mean	Std. Dev	MCSE
{LR:FOREIGN} < 0	0.9758	0.1538	0.0009
{LR:CR} > 0	1.0000	0.0000	0.0000
{LR:SIZE} > 0	0.7594	0.4275	0.0025
{LR:EQUITY} < 0	0.9979	0.0454	0.0003
{LR:IR} < 0	0.7167	0.4506	0.0026
{LR:SMR} < 0	0.9042	0.2943	0.0017
{LR:LDR} > 0	1.0000	0.0000	0.0000
{LR:ROA} > 0	0.6901	0.4625	0.0027
{LR:GDP} > 0	0.9999	0.0082	0.0000
{LR:DR} > 0	0.9305	0.2544	0.0015
{LR:NIM} > 0	0.5356	0.4987	0.0029

Source: Calculations of the author

Another result is that credit risk (CR) has a positive effect on liquidity risk (LR) with 100% probability. This result is consistent with the author's expectation and the results of studies by Delécha et al. (2012); Phan Thi My Hanh and Tong Lam Vy (2019). This explains why commercial banks in Vietnam focus mainly on lending and have a high NPL ratio; The higher the level of bad debt, the better the commercial bank's performance. In other words, if provisioning increases, profits of commercial banks will also increase. Commercial banks often increase lending while reducing the proportion of holding high-liquid assets to achieve profit goals. Therefore, when credit risk increases, there is a risk that commercial banks will run out of liquidity. The research results also show how closely the risks in commercial banks are related, showing that the State Bank must act quickly to protect commercial banks when they are in danger. If a commercial bank is in jeopardy, depositors will suffer significant losses, which will reduce public confidence in the commercial banking sector and increase the likelihood of a collapse of the financial system.

Besides, the loan/deposit ratio (LDR) is positively related to liquidity risk (LR) with the probability of 100%. This result is in line with the author's expectation and is consistent with the research results of Vu Thi Hong (2015); Bonfim and Kim (3). According to Golin (7), a larger ratio indicates that commercial banks are lending more than available capital. Therefore, if commercial banks lend too much while dealing with liquidity risk, it will be difficult to mobilize cheap capital, increasing liquidity risk. The author believes that when this ratio is low, commercial banks can easily mobilize capital from many different sources, including the commercial interbank market, issuing valuable papers, etc. increase their liquidity. Commercial banks allocate deposits to a certain extent between loans, investments and liquid assets in the market. Therefore, commercial banks will limit lending to liquid assets when this ratio is high, which indicates a high loan ratio, reducing the liquidity of commercial banks.

Bayesian regression results show that economic growth (GDP) positively affects liquidity risk (LR), the level of impact is very obvious when the probability is more than 99%. This result is contrary to the author's expectation but is consistent with the Bunda and Desquilbet (2003); Vodova (9); Cucinelli (2013). Vietnam's economy still has many risks, loans are not well secured, bad debt ratio is still high, it is difficult for commercial banks to recover debts. Therefore, the liquidity of commercial banks is also significantly affected. In addition, because commercial banks are the main source of capital for businesses during times of rapid economic expansion, they often cut their current assets while increasing lending, which increases the risk liquidity risk.

## 5 Conclusion and Policy Implications

From the research results and the reality of Vietnam's economy, the article argues that increasing the foreign ownership ratio to reduce liquidity risk in Vietnamese commercial banks is completely grounded in the integration period and developed as it is today. In credit institutions, the "rich" and "poor" banks have a clear division, so the state banks as well as the Government need to have a specific roadmap for each different group of banks in the banking system.

For groups of banks that operate inefficiently and have a high level of risk, the maximum ceiling to consider on the foreign ownership ratio can be as high as 100%. The reason is due to:

**Firstly**, the bank is a tool to help the state bank manage the currency in the economy, so if it operates inefficiently, it needs to be restructured comprehensively. When banks are weak, risks to the whole industry may occur due to the domino effect and crowd psychology that will cause people to go to bank branches to withdraw money to choose a safer investment channel. When the liquidity in the whole system is not enough, the crisis will happen like 2008.

**Secondly**, the self-restructuring resources of this group are almost non-existent because it is difficult to find a strategic partner with a maximum ownership level of only 30%, not enough 36% to have enough power to veto ineffective policies as well as to control ineffective policies like 51% to have the right to dominate the bank. Therefore, the 30% level is not really effective.

**Thirdly**, foreign investors when investing in a risky market like Vietnam are quite afraid as well as when banks operate inefficiently, investors have to choose other partners with a high level of stability and lower risk. Therefore, the foreign ownership rate at 30% will not attract strategic partners for this group of banks.

**Fourthly**, according to Clause 2, Article 149 of the Law on Credit Institutions 2010, it is stipulated: "The State Bank has the right to request the owner to increase capital, formulate and implement a restructuring plan or force a merger ..., consolidation or acquisition for a specially controlled credit institution, if the owner is unable or unable to carry out the capital increase". Therefore, the 100% ceiling is suitable for this group of banks.

For a group of banks with normal operations, the foreign ownership ceiling should be carefully considered between political and economic goals. The maximum political goal should be only 49% to avoid being taken over and dominate the entire financial market,

as well as creating autonomy for the economy to avoid being too dependent on foreign countries. The economic objective is to use foreign capital to improve equipment, modern technology, information security regime and improve human resources. In addition, the capital increase also helps banks complete the race to meet Basel II standards according to the set schedule.

For state-owned commercial banks, as the leading role in leading the entire banking system, the foreign ownership level of 0% is quite reasonable and does not need to be adjusted.

Besides, the research results show that liquidity risk and credit risk have a positive impact with a very large intercept. That is, liquidity is greatly affected by the debt collection ability of banks. Banks need to have policies to manage and handle bad debts as well as improve credit quality. To prevent problem debt, banks must improve the quality of inspection and supervision before, during and after lending.

The research results also show that the equity ratio has a negative impact on the bank's liquidity risk, that is, the higher the equity, the lower the liquidity risk. When the equity ratio increases, the bank will be less dependent on mobilized funds, reducing liquidity pressure. When equity increases, in addition to meeting Basel II standards, it also ensures the liquidity of banks for due deposits.

In addition, the loan/deposit ratio has a positive impact on liquidity risk. Most commercial banks in Vietnam only focus on lending mainly when the loan ratio accounts for more than 70% compared to other products and services at the bank, so the risk is quite large when bad debts increase. This means that the liquidity of the bank is vulnerable to serious damage. On the other hand, banks make profits based on NIM mainly without diversifying their own income, making their dependence on lending rates even higher. Therefore, in order to reduce liquidity risk, the bank must reduce the lending ratio, which means that the bank must implement policies to diversify income as well as use mobilized capital effectively.

Finally, the research results show that economic growth has a positive impact on liquidity risk. To minimize risks as well as attract foreign investors, first of all, information on the market needs to be transparent. In addition, the State Bank should have policies to help commercial banks manage bad debts and deal with problem debts. The bigger the economic growth, the more developed the economy and then the liquidity risk will be reduced.


## References

- Aspachs, O., Nier, E.W., Tiesset, M.: Liquidity, banking regulation and the macroeconomy (2005). Available at SSRN 673883
- Athanasoglou, P., Delis, M., Staikouras, C.: Determinants of bank profitability in the South Eastern European region (2006). MPRA Paper No. 10274
- Bonfim, D., Kim, M.: Liquidity risk in banking: is there herding. In: European Banking Center Discussion Paper, vol. 24, pp. 1–31 (2012)
- Demirgüç-Kunt, A., Huizinga, H.: Determinants of commercial bank interest margins and profitability: some international evidence. *World Bank Econ. Rev.* **13**(2), 379–408 (1999)
- Dinger, V.: Do foreign-owned banks affect banking system liquidity risk? *J. Comp. Econ.* **37**(4), 647–657 (2009)

- Duttweiler, R.: The meaning of liquidity risk. Chapter 1, 10–11 (2009)
- Golin, J.: The Bank Credit Analysis Handbook: A Guide for Analysts. In: Bankers and Investors. Wiley, Hoboken (2001)
- Terrell, H.S.: The role of foreign banks in domestic banking markets. In: Federal Reserve Bank of San Francisco Proceedings, pp. 297–304 (1984)
- Vodova, P.: Liquidity of Czech commercial banks and its determinants. *Int. J. Math. Model. Meth. Appl. Sci.* **5**(6), 1060–1067 (2011)
- Von Hagen, J., Ho, T.K.: Money market pressure and the determinants of banking crises. *J. Money Credit Bank* **39**(5), 1037–1066 (2007)
- Nguyen, T.X.H., Pham, T.H., Dao, T.N., Nguyen, T.N., Tran, T.K.N.: The impact of foreign ownership and management on firm performance in Vietnam. *J. Asian Finance Econ. Bus.* **7**(9), 409–418 (2020)
- Kusi, B.A., Kriese, M., Nabieu, G.A.A., Agbloyor, E.K.: Bank ownership types and liquidity creation: evidence from Ghana. *J. Afr. Bus.* **23**, 568–586 (2021)
- Le, T.: Can foreign ownership reduce bank risk? *Evid. Vietnam. SSRN Electr. J.* (13) (2021)
- Al-Harbi, A.: Determinates of Islamic banks liquidity. *J. Islamic Account. Bus. Res.* **11**(8), 1619–1632 (2020)



# Impacts of Financial Development on Vietnamese Commercial Banks' Lending Mechanisms of Monetary Policy Pass-Through: Bayesian Analysis

Thi Thu Hong Dinh<sup>2</sup>, Thanh Phuc Nguyen<sup>1</sup> , and Ngoc Tho Tran<sup>2</sup>

<sup>1</sup> Faculty of Finance and Banking, Van Lang University, 69/68 Dang Thuy Tram Street, Ward 13, Binh Thanh District, Ho Chi Minh City, Vietnam  
phuc.nt@vlu.edu.vn

<sup>2</sup> School of Finance, University of Economics Ho Chi Minh City, 59C Nguyen Dinh Chieu Street, Vo Thi Sau Ward, District 3, Ho Chi Minh City, Vietnam  
{hongtcdn, thotcdn}@ueh.edu.vn

**Abstract.** The research focuses on the driving role of financial development (represented by financial institution development) in the responses of bank loan supply to monetary policy shocks in Vietnamese commercial banks for the period of 2007–2019. The results from Bayesian analysis indicate that there is strong evidence for the presence of a bank lending channel of monetary policy transmission, which is in line with previous research on other developing countries. This might underline the bank-based economy of Vietnam, which is best suited to the transmission of diverse instruments of monetary policy via banks' granted loan supply. Furthermore, the greater the progress of financial development, the weaker the bank lending channel through which monetary policy can pass-through. This can be explained by the reduction in loanable funds of commercial banks, which may be replaced with external financing sources originating from the progress of financial markets or potential banking innovations. These findings remain qualitatively identical across indexes of financial development (especially an aggregate financial development indicator through principle component analysis) and a broad palette of monetary policy instruments. Given these findings, policy-makers could take into account the role of financial development when a banking system-based economy inevitably becomes more mature with diverse instruments for the need of financing and investment.

**Keywords:** Financial development · Bank lending mechanism · Monetary policy pass-through · Bayesian analysis · Vietnam

## 1 Introduction

A significant role that banks play in economic growth is based on the fact that they are a major financing source for businesses in many developing countries (Beck et al. 2000). Numerous research papers over the last decade have highlighted the importance

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 588–611, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_42](https://doi.org/10.1007/978-3-031-35763-3_42)

of monetary policy in determining bank loan supply (Kashyap and Stein 1995). There is a critical impact of monetary policy tightening (loosening) through an increase (a decrease) in interest rates on bank loan supply. This relationship displays a different aspect of the monetary policy transmission channel, the so-called bank lending mechanism of monetary policy pass-through (Altunbaş et al. 2002; Bashir et al. 2020). Although the pass-through of monetary policy through interest rates as a traditional channel remains intact, the credit channel's relevance in amplifying the impact of monetary policy has gained recognition (Peek and Rosengren 2010). Since the 2008 financial crisis, there has been increasing interest in exploring the function of bank lending in the transmission mechanism of monetary policy. This is covered in the bank's lending mechanism. The bank lending mechanism, as argued by (Bernanke and Blinder 1988), anticipates a fall in the aggregate volume of credit granted by banks in response to monetary policy tightening. Conversely, loosening monetary policy can encourage lending by increasing the availability of loanable funds for banks.

Many driving factors can explain why this channel has a distinct influence in different countries. The research looks at the relationship between bank individual factors and the bank lending channel (Gambacorta 2005; Leroy 2014). These studies offer evidence that bank features or balance-sheet strengths can influence responses of granted lending to changes in monetary policy. Due to severe agency costs in the deposit-based market, banks with weak balance sheet items face more difficulty raising uninsured funds, making their lending behavior more responsive to monetary policy shocks than banks with solid balance sheets (Kishan and Opiela 2006). Larger, more liquid, and well capitalized banks, for example, are more likely to shield their granted lending from monetary policy restrictions and are less vulnerable to monetary shocks.

Aside from bank-specific characteristics, there is broad agreement that financial sector transformations, such as deregulation in finance, innovative operations, and a competitive environment, lead to fundamental shifts in credit and bank markets, thereby influencing the potency of monetary policy transmission, primarily through the bank lending channel (Aysun and Hepp 2011; Boivin et al. 2010; Perera et al. 2014). Based on the relatively weak financial structures and high degree of financial friction of developing economies, the credit mechanism, rather than the interest rate mechanism, is viewed as the primary conduit of monetary policy pass-through (Mishra et al. 2016). To improve the operation of the pricing mechanism in resource allocation, a critical feature for developing economies' monetary policy framework transformation is to reduce financial frictions through financial development, hence reducing the strength of the credit channel (Zhan et al. 2021).

Our empirical analysis is conducted on the basis of a sample of 30 commercial banks in Vietnam over the period of 2007 to 2019. The analysis is illustrated by employing Bayesian analysis for bank-level panel data. By not depending on p-values of estimation coefficients, this approach can promote more reliable estimations than traditional dynamic panel estimation that is widely used in previous studies. Our findings indicate that the response of bank loan supply to monetary policy pass-through is statistically evidenced irrespective of monetary policy tools in each model; in other words, a bank lending channel through which monetary policy can be transmitted exists for the case of Vietnam. This finding confirms the existence of a bank lending mechanism in

emerging countries where the economy's development is mainly based on the banks' financing sources. In contrast to developed economies, empirical findings from emerging economies yield more conclusive results regarding the presence of the bank lending channel. Deposits are the principal source of bank funding in countries where the banking system is the primary financial channel (Freedman and Click 2006). In addition, as the degree of financial development increases, the monetary policy pass-through via the bank lending mechanism decreases. This means that the progress of the financial market through the development of the banking system makes the monetary policy ineffective. These results are robust to a variety of instruments for financial development and monetary policy.

This research contributes to the existing studies on the bank lending literature and the role of financial development in affecting the lending channel as follows. First, the dependence of monetary transmission via bank loan supply on financial development has been scarcely investigated. Although several studies have examined the linkage between financial development and the specific characteristics of monetary policy (Krause and Rioja 2006; Carranza et al. 2010), to the best of our understanding, very limited studies have explored the impact of financial development on the lending of commercial banks. Second, different from previous studies widely using the generalized method of moments for a dynamic panel data, we employ Bayesian analysis to provide reliable findings based on prior and posterior distribution. This approach can make inference findings more general. Furthermore, we use principal component analysis to facilitate a comprehensive picture concerning financial development, which remains scarce in prior research. Third, among research on the role of financial development in shaping the transmission of monetary policy in emerging countries, there is no research in Vietnam characterized by different features.

Vietnam offers a favorable and unique context to investigate the bank lending mechanism and the dependence of this mechanism on other factors (i.e., financial development among others). On the one hand, to grow, the Vietnamese economy is strongly reliant on the banking market, specifically bank lending. This may result in more pronounced monetary transmission via the banking channel (Saxegaard 2006). On the other hand, unlike previous studies, particularly those in developed countries that use the short-term interest rate as a single monetary policy indicator, this research uses a variety of monetary tools to assess the stance of monetary policy, such as the refinance rate, rediscount rate, and lending rate.

From this introduction, the remainder of this research is structured as follows. Section 2 summarizes the relevant literature relating to the bank lending mechanism and the dependence of this mechanism on other factors. Section 3 describes our empirical research design. The estimation results are reported and discussed in Sect. 4. Finally, Sect. 5 concludes the paper, followed by policy and research implications.



## 2 Literature Review

### 2.1 The Bank Lending Mechanism

Through a variety of mechanisms, including interest rates, exchange rates, other asset prices, and credit, monetary policy can have an impact on economic activity. One of the two distinct processes of the credit mechanism is the bank lending mechanism. It comes from the financial markets' limitations and depends on the incompatibility of bank loans and privately issued debt. The bank lending channel, as opposed to the conventional interest channel, emphasizes the unique role that banks play in the financial system as well as the close relationship between bank deposits (loanable funds) and loan supply (Bernanke and Blinder 1988). Bean et al. (2002) discuss the typical interest rate-based channel for monetary policy transmission, indicating that long-term interest rates is affected by changes in monetary policy. This impact through shifts of interest rates can drive a variety of relative prices in the economy and thus future consumption and investment associated with the present values. In contrast, Bernanke and Blinder (1988) suggest bank lending channel based on the incomplete substitutability of bank loans and bonds. To demonstrate the existence of the bank lending channel, the tightening of monetary policy should increase the opportunity cost of keeping deposits (loanable funds), resulting in a decrease in the availability of financing sources (bank loans). When financial institutions are unable to cope with the tightening of monetary policy (increase in interest rates by the central bank) and fall short of liquidity in terms of providing bank loans to borrowers due to an inability to access additional sources of funds, this inability has the effect of slowing the economy. In other words, the monetary policy restrictions can diminish reserves and, consequently, deposits, which make up the supply of loanable funds, according to the bank lending channel. If banks encounter difficulties issuing uninsured liabilities to offset a decrease in loanable funds, they will be forced to cut their loan portfolio (Bashir et al. 2020). This perspective assumes that bank loans are driven by policy-induced quantitative changes in deposits.

The post-2008 financial crisis has underlined the significance of financial intermediary functions in the transmission of monetary policy. The role of these intermediaries as loan providers is crucial to comprehending the effects of monetary policy on the economy. There is renewed interest in economic research and among practitioners in examining the bank lending channel as a monetary policy transmission mechanism (Altunbas et al. 2009, 2010; Brissimis and Delis 2009; Gambacorta and Marques-Ibanez 2011; Isakova 2008; Opiela 2008). According to this theory, monetary policy shocks may generate a change in banks' loan supply because it allows banks to get favorable access to loanable funds (Bernanke and Blinder 1988).

Shifts in banks' loan supply may exert a considerable impact on economic growth, as they either restrict or expand enterprises' bank-based financing sources. Understanding how monetary policy influences credit supply is therefore crucial for central banks and policymakers seeking to prompt economic growth. However, the majority of examinations are conducted in industrialized economies, particularly the United States and Europe, and demonstrate that the importance of this channel can vary across countries and regression periods (Altunbas et al. 2010; Brissimis and Delis 2009; Gambacorta and

Marques-Ibanez 2011; Isakova 2008). In contrast to industrial countries, fewer studies examine the bank lending channel in developing ones, but these studies are more conclusive about the existence of a bank lending mechanism (Amidu and Wolfe 2013). However, the intensity of this mechanism can differ between economies. Table 1 shows a summary of several pathways through which the monetary policy transmission through bank lending channel may be found, which potentially underlines the differences between interest rate channel and bank lending channel.

## 2.2 Impact of Financial Development on Monetary Policy Transmission via Bank Lending Mechanism

It is well established that the response of loan issuance to monetary impulses differs based on bank individual traits (Altunbas et al. 2010; Kashyap and Stein 1995; Kishan and Opiela 2006; Peek and Rosengren 2010). Due to their restricted ability to generate uninsured sources of funding, larger, more liquid, well capitalized, and riskier banks are more likely to buffer their lending from monetary restrictions. Furthermore, monetary policy becomes less effective in poorly competitive banking systems, primarily because larger banks with superior access to financing sources exist in these areas (Adams and Amel 2005, 2011; Olivero et al. 2011). However, according to the arguments of Amidu and Wolfe (2013) and Dang and Nguyen (2020), banks with a high strength of specific balance sheet items do not necessarily experience lower costs of external financing or get favorable access to non-deposit external financing, thus adding to the contradictory arguments of numerous prior studies.

Due to the fact that the existence and impact of the bank lending mechanism are mainly attributed to the imperfection of the financial market and the level of financial friction in the economy, any factors stemming from that may influence the monetary transmission mechanism, attracting considerable research on the bank lending mechanism from the standpoint of financial development. To be specific, according to Bashir et al. (2020), credit market imperfections could hold a unique function for financial intermediaries. These imperfections of the credit market may have a tendency to influence the bank's response to different monetary policies based on their natural capacity to raise external financing. That is, banks can protect their lending capacity against unfavorable shocks via monetary policy transmission. Consequently, in addition to the strength of bank-specific characteristics, numerous studies have examined the banking market features. In this vein, Adams and Amel (2005) discover that banks in banking markets with greater competition can lower their lending supply in response to monetary policy shocks. Gambacorta and Marques-Ibanez (2011) indicate that financial innovations and shifts in the business models of banks have a significant effect on the function of the bank lending channel as a monetary policy transmission mechanism. Moreover, Cantero-Saiz et al. (2014) found that banks operating in economies with a high level of sovereign risk are more influenced by tightening monetary policy.

After the 2008 financial crisis, there has been an increasing concern regarding the impact that financial development can have on financial intermediaries, their balance sheets, and their ability to extend credit. The financial crisis could induce a severe deterioration in the stability of the financial markets and banking system, causing researchers and policy-makers to be concerned about the changes in financial development. Studies

**Table 1.** Potential pathways for the impact of monetary policy on bank lending

Main mechanism	Types of loanable funds	Detailed pathway descriptions	Sources of pathway explanation
Changes in monetary policy → Loanable funds → Loan supply	The link between bank deposit and loan supply	Changes in monetary policy → Required reserves → Volume of deposits → Credit creation through the multiple impact → Bank lending	Friedman and Schwartz (1965) and Bernanke and Blinder (1988)
		Changes in monetary policy → The yields of bank deposit (compared to other assets) → Households' willingness to hold bank deposits → Bank lending	Kishan and Opiela (2000)
	The link between bank non-deposit sources of funding and loan supply (granted credit is not limited to deposits)	Changes in monetary policy → Changes in financial frictions → Changes in assets available to lend → Bank lending	Gibson (1997)
		Changes in monetary policy → Changes in overall risk portfolio of banks → Changes in credit standards → Bank lending	Maddaloni and Peydró (2011)
		Changes in monetary policy → Changes in risk perceptions of banks and bank balance sheets → Changes in cost of market funding → Changes in access to funding from financial markets → Bank lending	Disyatat (2011), Gambacorta and Marques-Ibanez (2011), and Cantero-Saiz et al. (2014)

Source: Author's compilation from previous research

from developing economies indicate that an underdeveloped market reduces the flexibility of bank credit in response to monetary restrictions because banks can experience a harder time obtaining loanable funds from alternative sources (Hou and Wang 2013).

Using panel data for Thailand and four other countries from 1999 to 2011, Lerskullawat (2017) concludes that endogenous and exogenous variables have distinct effects on the strength of monetary policy, with an increase in the bank's external capital market weakening the lending mechanism. Similar to studies conducted in industrialized economies, Olivero et al. (2009) employ a sample of Latin American economies and find that the credit mechanism can reduce the strength of monetary policy as bank competition decreases. More recently, Zhan et al. (2021) employ Chinese bank-level data for the period of 2010–2018 and show that the monetary policy pass-through via the bank lending mechanism is nonsignificantly affected by the progress of the money market.

Given that the bank lending mechanism works through the financial system, the degree of financial development can significantly influence its potency. In a reasonably developed financial system, banks have more exposure to the financial markets to insulate themselves from monetary shocks. Hence, the impact of the bank lending mechanism is able to be diminished. In these financial systems, banks may continue to be a substantial external financing source for non-financial enterprises, but the importance of financial markets is growing (Rybczynski 1997). As new market participants arise and new risk-trading instruments, such as derivatives, are developed, the traditional function of banks as collectors of deposits to provide lending sources is diminishing. Financial innovation may offer a new form of intermediation in which banks originate, repackage, and sell their loans to the financial markets. This securitization-based methodology appears to lessen the impact of monetary policy changes on loan supply (Altunbas et al. 2009). Ferreira (2010) indicate that the development effects of the banking sector and capital market on the less response of bank loan supply to monetary policy are evidenced on a sample of EU banks. Lerskullawat (2017) explains this link through a greater degree of financial progress determined by capital market advancements and the size of banks, which leads to the extension of the external funding potential of such banks. Therefore, it may be observed that financing sources can originate from the progress of financial markets, which helps banks to insulate the impact of monetary policy changes.

One should note that the direction of the financial system may change the way monetary policy can be transmitted via bank loan supply. Generally, the bank lending mechanism is more significant in countries with a greater reliance on bank financing than in economies with a market-based financial system (Brissimis and Delis 2009). Not only are banks less susceptible to monetary shocks, but enterprises that utilize a broader diversity of financing sources can also achieve some level of resilience. In this regard, Iturriaga (2000) finds that enterprises in economies with more market-oriented financial systems are less susceptible to changes in monetary policy than those in systems that rely more on bank financing.

In conclusion, the bank lending mechanism is likely to be more significant in developing economies. However, few studies make comparison for the case of various growing economies, and very scarce research examines the impact of financial development on this channel. In order to address this gap in the existing research, we will undertake an empirical analysis of this topic in the next section.

### 3 Methodology and Data

#### 3.1 Model Specification

Based on Kashyap and Stein (1995) and Cantero-Saiz et al. (2014), we propose the following estimation model to test the existence of a bank lending mechanism and the dependence of this channel on financial development as follows:

$$\Delta loan_{i,t} = \beta_0 + \beta_1 \Delta loan_{i,t-1} + \beta_2 \Delta MP_t + \beta_3 FD_t + \beta_4 \Delta MP_t * FD_t + \beta_5 BC_{i,t} + \beta_6 STATE_{i,t} + \beta_7 CRISIS_t + \beta_8 GDPG_t + \varepsilon_{i,t} \quad (1)$$

The dependent variable,  $\Delta loan_{i,t}$ , captures the growth rate in loan supply for bank  $i$  in year  $t$  compared to year  $t-1$ . This proxy has been generally employed in the bank lending mechanism literature (Gambacorta and Marques-Ibanez 2011; Jimborean 2009; Olivero et al. 2011). Similar to prior research, the 1-period lagged form of this variable is integrated as an independent variable to reflect the persistence effects of bank loan supply and the dynamic nature of the estimated model.

To reflect the diverse nature of monetary policy instruments in Vietnam's monetary framework, multiple interest rate-based vehicles for monetary policy stance ( $\Delta MP_t$ ) are used, including the lending rate ( $\Delta LENDR_t$ ), the short-term interbank interest rate ( $\Delta INTERBANK_t$ ), the refinance rate ( $\Delta REFINR_t$ ), and the rediscount rate ( $\Delta REDISR_t$ ). These policy interest rates, taking the first difference, are employed to represent the monetary policy regime in which the first-difference interest rate increases, implying the tightening regime followed by the reduction in the growth of loans and vice versa (Dang and Nguyen 2020; Sáiz et al. 2018).

$FD_t$  is the level of financial development, which consists of several proxies as follows. Because there are no widely-employed proxies for financial development that reflect its multifaceted nature, Levine (2002) and Beck and Levine (2002) provide an approach based on the principal component method (PCA). To capture the multidimensional features of financial development and facilitate understandings more comprehensively, the first principal component of several characteristics of financial systems is used. On the basis of this methodology, PC1 is constructed by the first principal component of the following financial system characteristics: financial institution depth index (FIDI), financial institution access index (FAI), and financial institution efficiency index (FIEI). These sub-items are collected from the financial development index database of the International Monetary Fund (IMF). Besides, the aggregate financial institution index (FII) retrieved from the IMF is also employed<sup>1</sup>. These financial development variables are included in separate models to eliminate collinearity and to compare findings when sequentially entering other indexes into models.

<sup>1</sup> For more reference, see detailed information of all indexes in this link: <https://data.imf.org/?sk=f8032e80-b36c-43b1-ac26-493c5b1cd33b&sid=1480712464593>. We do not use the financial market index (relating to stock market capitalization) in the IMF database because the case of Vietnam is characterized by the fact that economic agents depend mainly on the financing sources from the banking system and the stock market remains under-developed compared to other economies.

To qualify the appropriateness of the PCA approach, in panel A of Table A1 in the Appendix section, the PC1 component can explain approximately 61.06% of the variation in the original data. Therefore, PC1 is employed to form the CPA-based financial development index. To confirm the appropriateness of the PC1 component, we continuously use a scoring threshold of 0.3 or higher as a determinant of factor score significance. In panel B of Table A1, all scores have the excess significance of this cut-off value, indicating the effective use of PC1 as an aggregate PCA-based financial development index.

Furthermore, we aim to observe the shifts of the bank lending mechanism depending on the financial development by using the interaction terms between different indicators of monetary policy and financial development ( $\Delta MP_t * FD_t$ ), proposed by Kashyap and Stein (1995). This interaction term aims to investigate the response of bank loan supply to monetary policy shocks depending on the level of financial development. In this regard, a significantly positive parameter  $\beta_4$  (which is a coefficient of our main interest) denotes the degree of financial development increasing while the monetary policy pass-through via the bank lending mechanism deteriorates.

$BC_{i,t}$  denote several bank-specific characteristics, consisting of a share of total equity over total assets (CAP); is natural logarithm of total assets (SIZE), the ratio of cash and deposits to total assets (LIQ), and the proportion of loan loss provision divided by gross loan (LLP). In the light of Olivero et al. (2011), Zhan et al. (2021), Cantero-Saiz et al. (2014), and Sanfilippo-Azofra et al. (2018), we use relative measures by normalizing these bank-specific variables with respect to their sample mean across all banks.

$STATE_{i,t}$  refers to a dummy variable for banks with state ownership, taking a value of 1 in the case of state-owned banks and 0 otherwise. Besides, in the period of 2011–2014, the lending expansion of most Vietnamese commercial banks has resulted in an asset-quality crisis and, thereby, Vietnam's banking crisis (Pham 2021). To control the impact of this crisis, we use the dummy variable,  $CRISIS_t$ , which takes the value of 1 for the 2011–2014 crisis period and 0 otherwise.  $GDPG_t$  is the growth rate of gross domestic product, which is introduced to distinguish the supply-side bank lending channel from the alternative demand-side interest rate channel (Yang and Shao 2016). This macroeconomic variable is used to control for the business cycle, having a tendency to affect the credit supply positively (Jimboean 2009).

The descriptive statistics and correlation amongst studied variables are reported in Table 2 and Table 3, respectively. As reported in Table 3, there is no concern about severe multicollinearity issues in the model equation, and independent variables that have high correlation values with other variables are treated in separate models to avoid spurious regression.

### 3.2 Regression Estimation Methods: Bayesian Analysis

This research use the Bayesian approach rather than dynamic panel estimation as widely employed in previous studies. The p-value from traditional regression measures the deviation between the data and the null hypothesis of interest, which often assumes there is no difference or impact. A Bayesian strategy permits the calibration of p-values by translating them into direct estimates of the evidence against the null hypothesis, known as Bayes factors (Held and Ott 2018). In contrast to the p-value, which only

**Table 2.** Descriptive statistics of variables

Variable	Descriptions	Mean	Std. Dev	Min	Max
$\Delta$ loan	The growth of granted loans	35.163	60.629	-49.160	295.140
$\Delta$ LENDR	The short-term lending rate	-0.230	2.760	-5.715	4.604
$\Delta$ REFINR	The refinancing rate	-0.120	2.580	-4.000	5.330
$\Delta$ REDISR	The rediscount rate	-0.092	2.771	-4.430	5.750
$\Delta$ INTERBANK	The short-term interbank rate	-0.238	2.918	-5.340	6.140
CAP	The ratio of bank's capital to total assets	10.388	5.133	4.275	26.621
SIZE	The natural logarithmic form of total assets	11.163	1.250	8.681	13.887
LIQ	The ratio of liquid assets to total assets	20.408	10.495	5.798	50.587
LLP	The loan loss provision divided by gross loan	0.642	0.694	0.000	3.900
FIAI	The financial institution access index	12.904	2.815	5.700	16.000
FIDI	The financial institution depth index	26.335	4.111	20.200	32.500
FIEI	The financial institution efficiency index	71.959	1.956	68.600	75.500
FII	The aggregate financial institution index	34.872	2.893	30.200	39.600
PC1	The financial development calculated based on principal component analysis	0.000	1.353	-2.205	2.306
STATE	State ownership	0.103	0.304	0.000	1.000
CRISIS	2011–2014 banking crisis in Vietnam	0.255	0.437	0.000	1.000
GDPG	The growth rate of gross domestic product	6.249	0.643	5.247	7.130

Note: For the purpose of comprehensive understanding, the descriptive statistics of bank-specific variables such as CAP, SIZE, LIQ, and LLP are shown in the standard form before the normalization  
Source: Author's calculation from using Stata 15.1 software

offers information about the probability that the null hypothesis is true, the Bayes factor captures both the null and alternative hypotheses directly. The Bayes factor evaluates the relative evidence in the gathered data on whether those data are well predicted by the null or alternative hypothesis (an effect of stated magnitude) (Halsey 2019). The

Table 3. Correlations between studied variables

	$\Delta loan$	$\Delta LEADR$	$\Delta REFINR$	$\Delta REDISR$	$\Delta INTERBANK$	CAP	SIZE	LIQ	LLP
$\Delta loan$	1								
$\Delta LEADR$	-0.083	1							
$\Delta REFINR$	-0.087	0.894 <sup>***</sup>	1						
$\Delta REDISR$	-0.091	0.91 <sup>***</sup>	0.998 <sup>***</sup>	1					
$\Delta INTERBANK$	-0.067	0.934 <sup>***</sup>	0.902 <sup>***</sup>	0.913 <sup>***</sup>	1				
CAP	0.031	0.018	0.005	0.009	0.037	1			
SIZE	-0.233 <sup>***</sup>	0.012	0.006	0.004	-0.017	-0.709 <sup>***</sup>	1		
LIQ	0.264 <sup>***</sup>	-0.043	-0.027	-0.026	-0.017	0.241 <sup>***</sup>	-0.348 <sup>***</sup>	1	
LLP	-0.176 <sup>***</sup>	-0.028	-0.012	-0.015	-0.022	-0.039	0.229 <sup>***</sup>	-0.352 <sup>***</sup>	1
FIAI	-0.432 <sup>***</sup>	0.103 <sup>*</sup>	0.072	0.071	0.042	-0.373 <sup>***</sup>	0.500 <sup>***</sup>	-0.525 <sup>***</sup>	0.221 <sup>***</sup>
FIDI	-0.154 <sup>**</sup>	0.061	0.069	0.057	0.019	-0.388 <sup>***</sup>	0.442 <sup>***</sup>	-0.436 <sup>***</sup>	0.185 <sup>***</sup>
FIEI	0.240 <sup>***</sup>	0.005	0.009	0	-0.024	-0.122 <sup>*</sup>	0.091	0.013	0.015
FII	-0.227 <sup>***</sup>	0.08	0.073	0.064	0.025	-0.406 <sup>***</sup>	0.485 <sup>***</sup>	-0.472 <sup>***</sup>	0.204 <sup>***</sup>
PC1	-0.209 <sup>***</sup>	0.079	0.07	0.062	0.022	-0.402 <sup>***</sup>	0.479 <sup>***</sup>	-0.459 <sup>***</sup>	0.200 <sup>***</sup>
STATE	-0.095	-0.013	0.002	0.002	-0.01	-0.269 <sup>***</sup>	0.541 <sup>***</sup>	-0.06	0.120 <sup>*</sup>
CRISIS	0.012	-0.116 <sup>*</sup>	-0.022	-0.022	-0.137 <sup>**</sup>	-0.136 <sup>**</sup>	0.089	0.004	-0.009
GDPG	0.138 <sup>**</sup>	-0.003	0.021	0.012	-0.009	-0.248 <sup>***</sup>	0.219 <sup>***</sup>	-0.171 <sup>***</sup>	0.056
FIAI		FIDI	FIEI	FII	PC1	STATE	CRISIS	GDPG	
FIAI	1								
FIDI	0.726 <sup>***</sup>	1							
FIEI	-0.086	0.479 <sup>***</sup>	1						
FII	0.828 <sup>***</sup>	0.980 <sup>***</sup>	0.426 <sup>***</sup>	1					
PC1	0.803 <sup>***</sup>	0.978 <sup>***</sup>	0.480 <sup>***</sup>	0.998 <sup>***</sup>	1				
STATE	-0.016	-0.009	0.006	-0.011	-0.011	1			
CRISIS	0.02	0.203 <sup>***</sup>	0.363 <sup>***</sup>	0.193 <sup>***</sup>	0.212 <sup>***</sup>	0.001	1		
GDPG	0.189 <sup>***</sup>	0.640 <sup>***</sup>	0.794 <sup>***</sup>	0.598 <sup>***</sup>	0.633 <sup>***</sup>	0.003	0.348 <sup>***</sup>	1	

Source: Author's calculation from using Stata 15.1 software



robustness of estimation model is a problem with the estimating technique. Due to the fact that the regression coefficients of variables in a model can change as the number of observations varies, the estimation results may alter the conclusions drawn. Due to the objective constraint on sample data for estimation by traditional estimation methods, the relationship between variables may be influenced in this research. To address this deficiency and further consolidate the estimators, the Bayesian technique is employed to confirm the model’s robustness.

The Bayesian analysis is approached from the conditional probability as follows:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2}$$

where  $p(A|B)$  denotes posterior probability, suggesting the need to obtain the probability that Hypothesis A holds true based on the collected data;  $p(B|A)$  refers to the appropriateness of the database (likelihood), representing the probability of collected data given the situation that Hypothesis A holds true;  $p(A)$  is prior probability, indicating that the probability that Hypothesis A holds true before collecting the data;  $p(B)$  stands for constant term, which is the probability of data;  $A$  and  $B$  are random vectors.

In Bayesian analysis, the parameters estimated from the regression model are random vectors. The Bayesian analysis defines the posterior distribution of these estimated parameters based on the mixture of the given data and the a priori distribution. Each posterior distribution displays the probability distribution of a certain parameter in the model. Therefore, the approximate inference findings with respect to estimated parameters in research specifications on the basis of posterior distributions become more general.

$$\text{Posterior distribution} \propto \text{Likelihood function} \times \text{Prior distribution.}$$

One should note that there is a need to identify the prior distribution and the sampling algorithm when applying the Bayesian-based approach.

- The prior distribution

A prior distribution can be categorised as informative or non-informative, which may be employed to yield a posterior distribution. Available information for the estimated parameters plays a pivotal role in Bayesian-based inference. The prior distribution includes information concerning previously observed parameters. If the larger data is sufficient, the prior distribution will exert minimal effect on the posterior distribution. On the contrary, in case the data are too small, the prior distribution will hold a dominance in the posterior distribution. The prior distributions of the estimated coefficients will be identified by normal distributions. In more details,

$$\beta_i \sim N(\hat{\beta}, \hat{\sigma}_\beta^2) \tag{3}$$

where  $\hat{\beta}$  denotes the estimated coefficient;  $\hat{\sigma}_\beta^2$  stands for the standard deviation of the estimated coefficients.

- The sampling algorithm

The Metropolis–Hastings approach, one of the MCMC methods, is widely employed. This method was originally proposed by Metropolis et al. (1953) and consequently developed by Hastings (1970) with a more efficient version. In this study, the Metropolis–Hastings sampling algorithm is employed to create an MCMC (Markov chain Monte Carlo) chain with a size of 25,000 and remove 2,500 at the burn-in stage. The sample size for MCMC analysis finally turns out to be 22,500.

### 3.3 Research Data

A total of 30 Vietnamese commercial joint stock banks are included in a research sample covering the period of 2007–2020. Annual bank-level data on the balance sheets and income statements is collected from the Bankscope database. Besides, to obtain the data accuracy, the audited financial reports from the websites of each bank are also used. In addition, banks involved in mergers or acquisitions during the study period are excluded. The final sample consists of 409 observations, which represents approximately 90 percent of the total assets of the Vietnamese banking system in any given year. Before using regression treatment, the winsorizing procedure is used to the micro-level variables at the 2.5% and 97.5% interval levels in order to isolate the possibly deteriorated effects of extreme values.

## 4 Empirical Results and Discussion

### 4.1 Baseline Regression Results

Table 4 reports findings on the existence of Vietnam's bank lending mechanism of monetary policy pass-through and the role of financial development in shaping the sensitivity of bank loan supply to monetary policy impulses. Columns 2–3, 4–5, 6–7, 8–9, and 10–11 capture a single indicator of monetary policy and a series of financial development indexes (i.e., FIAI, FIDI, FIEI, FII, and PC1, respectively). Each panel (A, B, C, and D) represent different single proxy of monetary policy such as the lending rate, rediscount rate, refinance rate, and interbank rate, respectively.

As displayed in the combined columns and panels of Table 4, the lagged form of a dependent variable such as the growth rate of bank loan supply reports a significantly positive coefficient through all columns of the financial development index, showing the persistence nature of bank loan growth in the continuous years. The significantly negative coefficients of monetary policy instruments such as interbank rate, refinancing rate, lending rate, and lending rate provide evidence that the bank lending channel of monetary policy transmission exists in the case of Vietnam as a typical emerging economy. This means that a tightening monetary policy (an increase in policy interest rate) leads to a reduction of bank loan supply and vice versa, which is suitable for the case of emerging countries in which banks' financing sources for economic agents are dominant. Therefore, these tools of monetary policy can be harmoniously used to navigate the economy towards stability in a given turbulent period. This highlights the effectiveness of diversified instruments of Vietnam's monetary policy in practice, which might be relatively different from other emerging countries.

**Table 4.** Monetary policy and financial development indexes: Bayesian analysis

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]
Panel A	ΔLENDR&FIAI		ΔLENDR&FFDI		ΔLENDR&FIEI		ΔLENDR&FHI		ΔLENDR&PCI	
MPR	-5.87	-2.07	-7.15	-10.83	-29.76	-45.96	-14.28	-21.37	-1.55	-2.88
FD	-4.27	-0.81	4.42	2.05	6.76	-1.11	4.11	-0.44	6.37	-1.14
MPR*FD	0.37	0.64	0.23	0.10	0.36	0.17	0.63	0.17	0.76	0.19
Lag.LG	-0.004	0.08	0.10	0.02	0.18	-0.04	0.12	-0.01	0.07	-0.01
CAP	2.74	4.04	3.65	2.40	4.90	1.99	4.51	2.26	3.56	2.27
SIZE	0.64	7.58	4.08	-2.77	10.91	+ 4.42	9.32	-3.50	3.52	-3.53
LIQ	0.40	-0.11	0.92	0.20	1.17	0.54	1.03	0.14	0.65	0.14
LLP	7.01	-0.05	13.96	-1.85	11.98	6.14	13.17	-1.52	5.51	-1.60
STATE	-1.39	-20.22	17.55	-20.21	18.14	-1.28	18.08	-1.01	-0.96	-20.39
CRISIS	-10.21	-21.38	0.88	-24.25	-1.90	-22.50	0.15	-23.59	-12.81	-24.36
GDP	15.27	5.59	-12.29	-26.57	1.90	-5.81	16.08	-16.67	0.19	-13.58
Panel B	ΔREDISR&FIAI		ΔREDISR&FFDI		ΔREDISR&FIEI		ΔREDISR&FHI		ΔREDISR&PCI	
MPR	-8.17	-11.50	-5.71	-10.27	-0.96	-47.44	-14.03	-22.62	-1.65	-2.62
FD	-4.38	-7.79	4.44	2.08	6.77	-1.16	4.06	-0.50	6.05	-1.48
MPR*FD	0.53	0.30	0.18	0.01	0.35	0.19	0.65	0.11	0.59	0.14
Lag.LG	-0.002	-0.09	0.08	0.02	0.18	-0.04	0.12	-0.01	0.07	-0.01
CAP	2.75	1.45	3.67	2.43	4.90	2.02	4.52	2.27	3.58	2.26
SIZE	0.76	-6.32	7.67	-2.57	11.13	-4.37	9.43	-3.30	3.70	-3.26
LIQ	0.41	-0.10	0.92	0.21	1.19	0.54	1.05	0.15	0.66	0.151
LLP	7.16	0.14	14.18	-1.71	12.26	6.13	13.11	-1.51	5.66	-1.45
STATE	-1.55	-20.67	17.82	-20.17	17.84	-1.30	18.15	-20.55	-1.20	-20.41
CRISIS	-9.79	-20.91	1.13	-23.60	-1.36	-22.14	0.45	-23.23	-11.92	-23.48

(continued)

Table 4. (continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	
	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	Mean	[95% Cred. Interval]	
GDP	15.28	5.73	-12.62	-26.82	5.29	-5.86	-1.59	-16.64	0.46	-13.38	14.17
Panel C	$\Delta$ REFIR&FIAI	$\Delta$ REFIR&FIDI	$\Delta$ REFIR&FIEI	$\Delta$ REFIR&FII	$\Delta$ REFIR&FIII	$\Delta$ REFIR&FIV	$\Delta$ REFIR&FVI	$\Delta$ REFIR&FVII	$\Delta$ REFIR&FVIII	$\Delta$ REFIR&FVII	$\Delta$ REFIR&FVIII
MPR	-7.79	-11.26	-4.22	-12.07	-32.76	-50.50	-12.62	-22.26	-1.80	-2.80	-0.80
FD	-4.37	-7.81	-0.90	2.09	6.85	-1.17	3.23	-0.46	6.16	-1.44	13.85
MPR*FD	0.50	0.25	0.75	0.11	0.42	0.20	0.34	0.05	0.67	0.15	1.19
Lag.LG	-0.003	-0.09	0.08	0.02	0.18	-0.04	0.08	-0.01	0.07	-0.01	0.16
CAP	2.75	1.46	4.05	2.44	4.91	2.02	3.55	2.29	3.59	2.28	4.90
SIZE	0.83	-6.16	7.83	-2.62	11.06	-4.41	3.71	-3.24	3.72	-3.35	10.74
LIQ	0.41	-0.10	0.93	0.21	1.20	0.05	0.65	0.15	0.66	0.15	1.17
LLP	7.11	0.10	14.07	-1.82	12.10	-0.98	5.63	-1.49	5.68	-1.41	12.72
STATE	-1.71	-21.00	17.44	-20.26	17.77	-20.59	-1.43	-20.66	-1.31	-20.46	17.97
CRISIS	-9.80	-20.96	1.27	-23.78	-1.66	-22.22	-11.88	-23.13	-11.96	-23.38	-0.57
GDP	15.30	5.56	24.99	-26.88	1.73	-5.81	-1.70	-16.57	0.57	-3.26	14.48
Panel D	$\Delta$ INTERBANK&FIAI	$\Delta$ INTERBANK&FIDI	$\Delta$ INTERBANK&FIEI	$\Delta$ INTERBANK&FII	$\Delta$ INTERBANK&FIII	$\Delta$ INTERBANK&FIV	$\Delta$ INTERBANK&FVI	$\Delta$ INTERBANK&FVII	$\Delta$ INTERBANK&FVIII	$\Delta$ INTERBANK&FVII	$\Delta$ INTERBANK&FVIII
MPR	-15.85	-21.99	-9.76	-15.56	-0.02	-83.35	-18.31	-35.42	-1.12	-1.73	-0.53
FD	-4.44	-7.89	-0.94	1.98	6.70	-1.18	3.00	-0.72	5.93	-1.56	13.43
MPR*FD	1.08	0.64	1.52	-0.03	0.54	0.10	0.49	0.03	0.61	0.15	1.05
Lag.LG	0.01	0.01	-0.08	0.02	0.19	-0.03	0.08	-0.005	0.07	-0.01	0.15
CAP	2.70	2.70	1.37	2.43	4.90	2.01	3.54	2.26	3.56	2.27	4.84
SIZE	0.41	0.41	-6.75	-2.78	10.89	-4.55	3.44	-3.52	3.63	-3.29	10.60
LIQ	0.43	0.42	-0.09	0.20	1.19	0.05	0.64	0.134	0.65	0.14	1.15
LLP	7.04	7.04	0.07	-1.93	12.22	-1.12	5.41	-1.67	5.57	-1.46	12.68
STATE	-0.54	-0.54	-19.69	-20.17	17.96	-20.46	-0.81	-20.29	-1.35	-20.63	17.89
CRISIS	-10.46	-10.46	-21.8	-24.10	-1.81	-22.47	-12.31	-23.63	-12.32	-23.77	-0.96
GDP	15.92	15.92	6.15	-26.08	2.67	-5.20	-0.48	-15.68	0.89	-12.93	14.57

Source: Author's calculation from using Stata 15.1 software

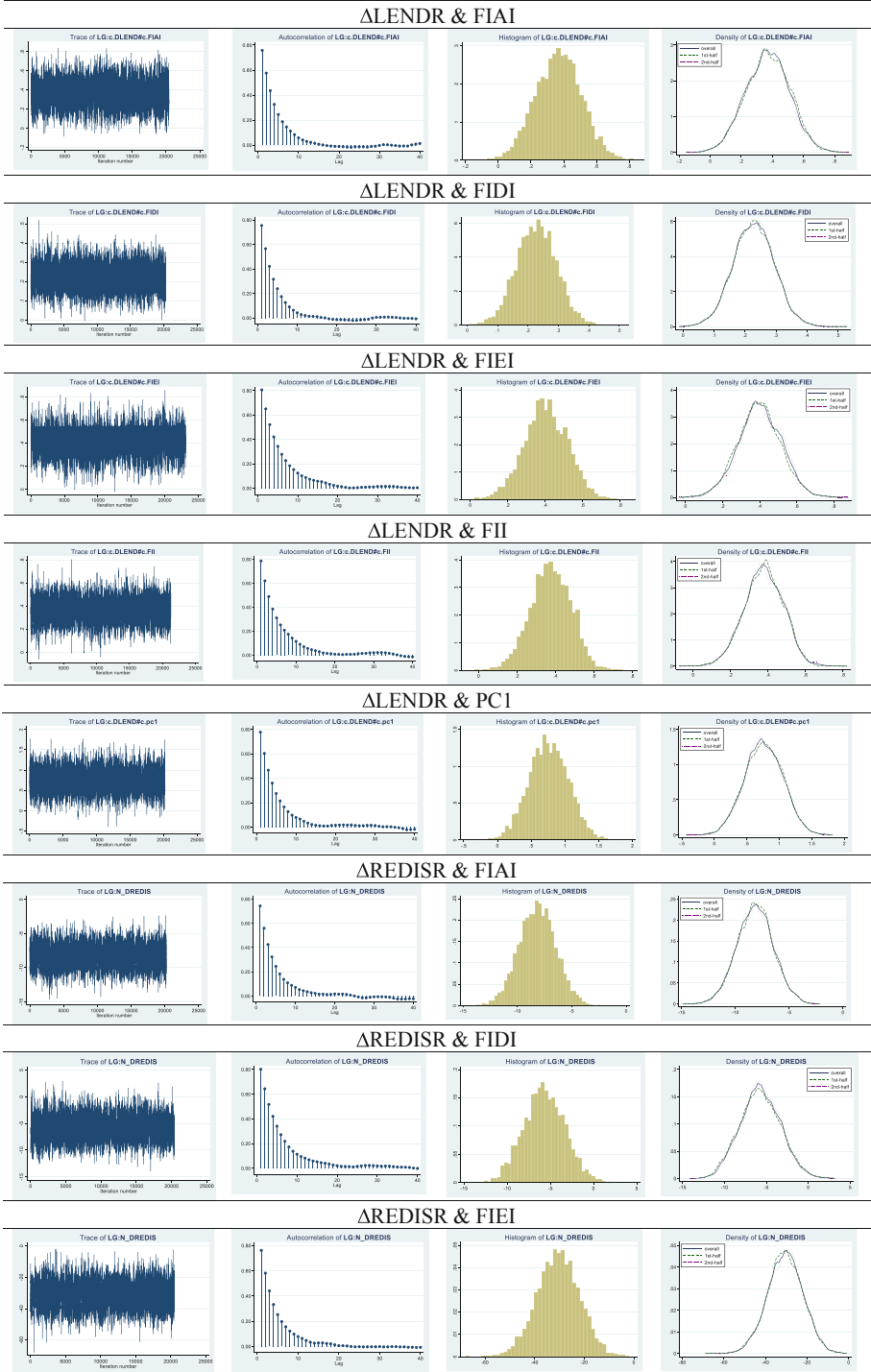
The Bayesian-based estimation coefficients of interaction terms between financial development indexes and monetary policy instruments all fall within the 95% confidence interval of their respective regression coefficients, implying that the value of this regression coefficient is in the positive value domain. The Bayesian-based estimators of the interaction term between all proxies of the financial development index and monetary policy instruments have been scarcely explored. Therefore, as financial development proceeds, the bank lending channel of monetary policy can deteriorate. More specifically, the progress of financial market through the development of banking institution can make the monetary policy transmission via bank lending mechanism more ineffective. This result confirms that the development of financial markets may have an impact on the way in which monetary policy can be transmitted via bank lending mechanisms. Similar to Lerskullawat (2017) who investigates the financial market development (involving the progress of the banking system and stock market), the banking sector development with respect to banking operations can result in a weaker impact of monetary policy through the lending mechanism. Besides, our research results are not in line with Sanfilippo-Azofra et al. (2018) who show that only after the financial crisis, as the financial system is more developed, the effectiveness of the bank lending mechanism through which monetary policy can be transmitted is evidenced.

The level of financial intermediation with respect to the financial institutions' size and liquidity may be influenced by banking system development (Gertler 1991). This can remove financial costs and stimulate the strength of the bank's balance sheet items. Therefore, financial development could reduce the pass-through strength of monetary policy via bank lending mechanisms. This may stem from more opportunities to reach externally financed sources, which banks can be exposed to (Altunbas et al. 2009; Ferreira 2010). Given this result, we agree with the theoretical perspective that bank loan supply can be replaced with external financing funds, which could insulate the impact of monetary policy impulses on loanable funds. It means that the failure of the external financing market may facilitate the strength of the bank lending channel of monetary policy pass-through (Zhan et al. 2021).

Turning to other control variables, we observe that through most models in which the diverse proxies of financial development index and monetary policy instruments, and other factors remain equal, bank capital (CAP) has a statistically positive effect on the bank loan supply, which is consistent with Kishan and Opiela (2006) who indicate that an increase in loan growth is attributed to an increase in the bank's capital. With regard to credit risk, the regression result shows that banks with a higher level of loan loss provision have a tendency to increase bank loan supply. This is contrary to the work of Altunbas et al. (2010) who propose that banks with higher credit risk tend to have lower growth rates of bank loan supply. This difference can be explained by the fact that when facing an increase in the share of loan loss provision, Vietnamese commercial banks can cover the reduction in revenue stream by providing more credit. This means that the sensitivity of bank loan supply is stronger for banks experiencing a high level of credit risk (Vo and Nguyen 2014).

#### **4.2 Converging Testing of MCMC Chain After Bayesian Analysis**

In the Fig. 1 below, the MCMC chain testing result for the regression coefficient on the interaction term of financial development and monetary policy indicates that the MCMC chain shows statistically converging features. For more details, the trace chart (the first one for each case of interaction term) displays that the MCMC chain has no specific trend and fluctuates around the mean value of the interaction term. Accordingly, the estimators illustrate the thick distribution into a horizontal line fluctuating around this mean value. The autocorrelation chart (the second one) displays a decreasing correlation to zero. The distribution chart (the third one) of the estimated coefficient of the interaction term follows a normal distribution. Further, the density functions of an MCMC half-chain, front, back, and overall, are nearly identical. Thus, the results of Bayesian analysis qualitatively all show the identical sign of the interaction term, lending support for research inference from the aforementioned findings of Sect. 4.1.



**Fig. 1. Robust diagnostics for MCMC convergence.** Source: Author’s illustration from using Stata 15.1 software

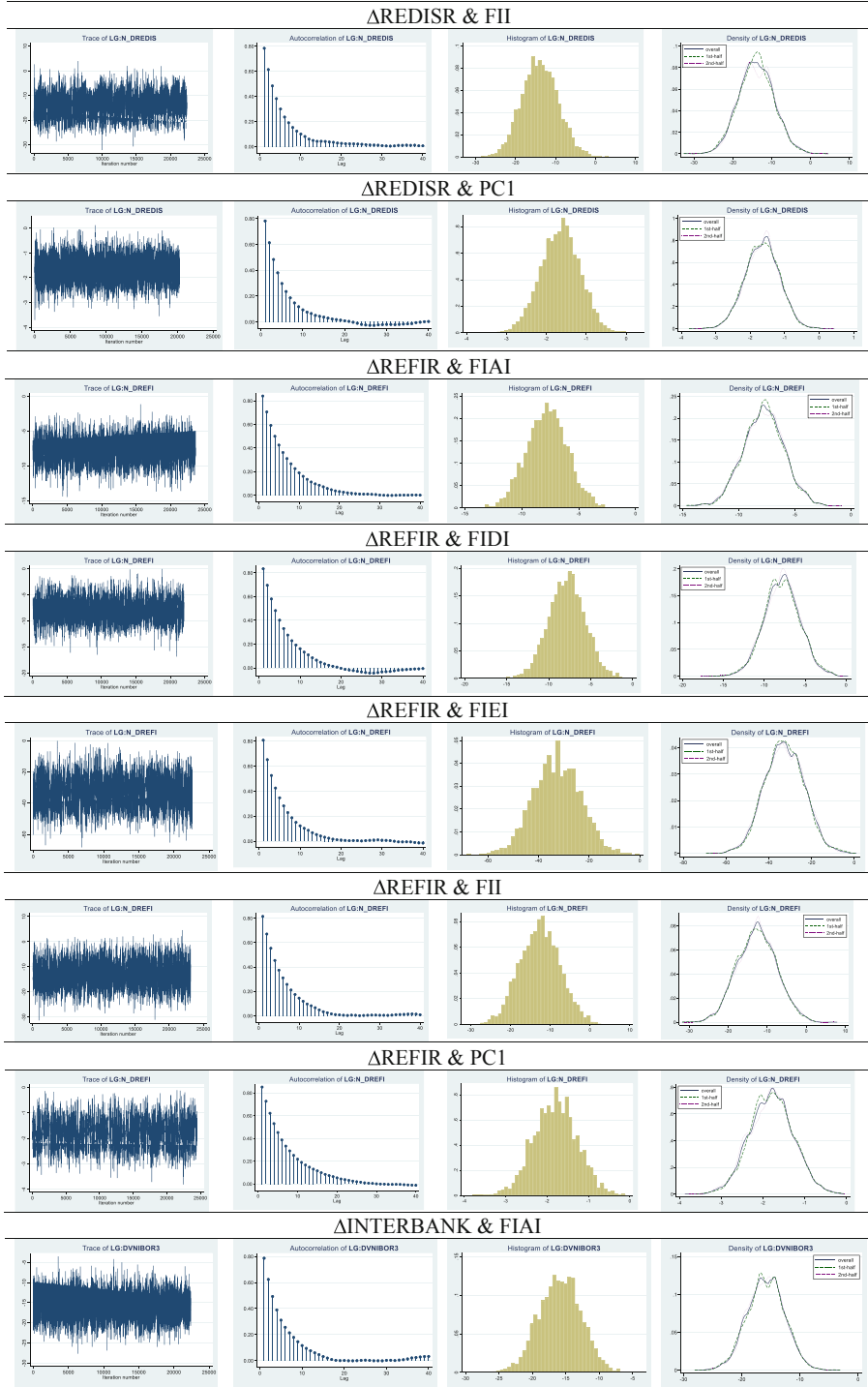


Fig. 1. (continued)



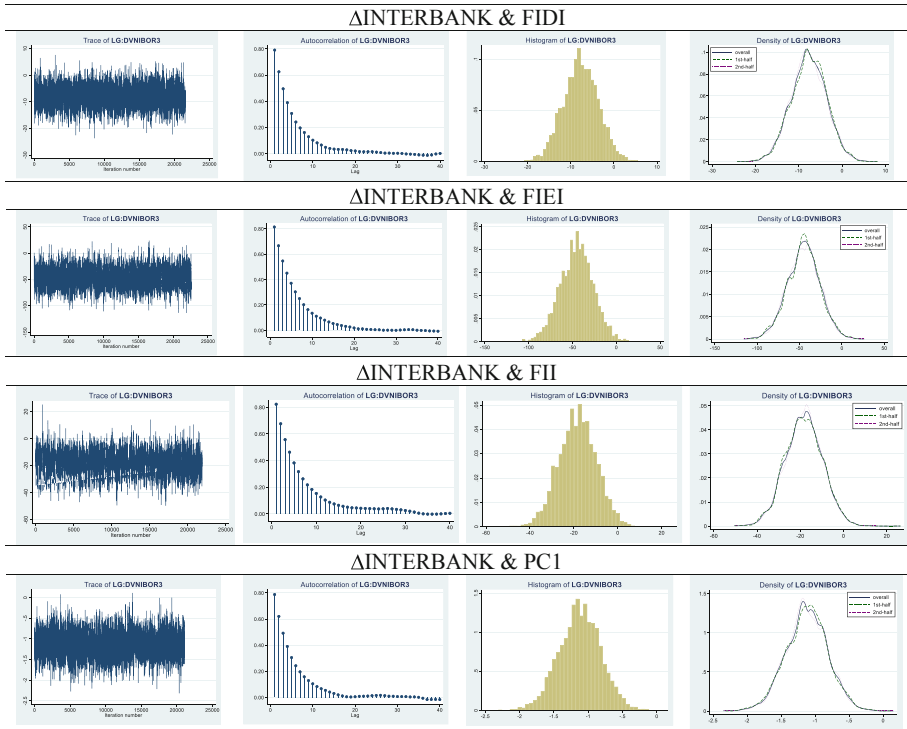


Fig. 1. (continued)

## 5 Conclusion

The lending channel is considered a critical conduit of monetary policy pass-through which can display the impact of monetary policy via bank loan supply on the economy (Mishkin 1996). Because of the driving role of the banking sector and financial market in bank loan supply, financial market development (i.e., the progress of the banking sector) can lead to significant impacts on the banking sector and credit market (Altunbas et al. 2009). This highlights the essential problem of how these shifts can influence the way in which monetary policy affects the economy via the bank lending mechanism. The design of monetary policy pass-through via bank lending channels can depend on the relative criticality of the financial market. This gives rise to the need to investigate the so-called bank lending mechanism of monetary policy transmission, which has received a great deal of attention in previous studies, possibly depending on the level of financial market development. As the bank lending mechanism works through the financial system, the aim of this research is to examine whether the potency of this mechanism can depend on the development of the financial system.

Employing an unbalanced sample of 30 commercial banks from 2007 to 2020 and the Bayesian analysis, we confirm the bank loan supply persistence and the existence of bank lending channel in Vietnam with updated bank-level data and normalized procedure for bank-specific characteristics. Furthermore, we find the bank loan supply in the case

of more developed financial systems is significantly and positively affected by monetary policy restrictions (negatively affected by monetary policy loosening). This means that the development of financial development can weaken the transmission of monetary policy via the bank lending mechanism. It is explained that financial development can lead to a fall in banks' dependence on loanable funding sources and stimulate the banks' exposure to external financing funds, making monetary policy more ineffective. The aforementioned results hold true when using all proxies of monetary policy and financial development indexes. Therefore, we can hold the view that financial development is considered as one of the critical drivers to affect the potency of monetary policy pass-through via bank lending channel.

The importance of monetary policy conduct as an instrument for macroeconomic policy has been established, and understanding the mechanism of transmission is a key to effective implementation of monetary policy. In this research, financial development is evidently an important factor in weakening the impact of monetary policy on bank loan supply. This suggests that the State Bank of Vietnam (SBV) should take into account the development of the financial systems when implementing their monetary policy to boost loan supply, especially for emerging countries in which the banking system remains an important financing source for firms and households and experiences continuous growth. This research could not avoid limitations, which gives rise to further investigation in further future research. We limit the monetary policy vehicles to the interest rate-based ones, which may be a representative case for all the means that monetary policy can transmit. Further research can extend this research by replacing the traditional interest rate with unconventional monetary policy such as open market operations or SBV purchases and foreign exchange reserves.

**Acknowledgements.** This research is funded by the University of Economics Ho Chi Minh City, Vietnam (UEH). This study is a main part of the doctoral thesis of Thanh Phuc Nguyen under the supervision of Tho Ngoc Tran at the UEH. The second author appreciates the financial support from Van Lang University.

## Appendix

**Table A1.** Principal component analysis to obtain financial development index (PC1)

Panel A: Cumulative explanation for variation in the original data				
Components	Eigenvalue	Difference	Proportion	Cumulative
PC1	2.0832	0.7524	0.6106	0.6106
PC2	1.0794	0.9906	0.3598	0.9704
PC3	0.0888		0.0296	1.0000
Panel B: Scoring significance values				
Variables	PC1	PC2	PC3	
FIAI	0.5936	-0.5476	0.5897	
FIDI	0.7224	0.0397	-0.6903	
FIEI	0.3547	0.8358	0.4192	

Source: Author's calculation from using Stata 15.1 software

## References

- Adams, R.M., Amel, D.F.: The effects of local banking market structure on the bank-lending channel of monetary policy (2005). Available at SSRN 716349
- Adams, R.M., Amel, D.F.: Market structure and the pass-through of the federal funds rate. *J. Bank. Finance* **35**(5), 1087–1096 (2011)
- Altunbaş, Y., Fazylov, O., Molyneux, P.: Evidence on the bank lending channel in Europe. *J. Bank. Finance* **26**(11), 2093–2110 (2002). [https://doi.org/10.1016/S0378-4266\(02\)00201-7](https://doi.org/10.1016/S0378-4266(02)00201-7)
- Altunbas, Y., Gambacorta, L., Marques-Ibanez, D.: Securitisation and the bank lending channel. *Eur. Econ. Rev.* **53**(8), 996–1009 (2009). <https://doi.org/10.1016/j.eurocorev.2009.03.004>
- Altunbas, Y., Gambacorta, L., Marques-Ibanez, D.: Bank risk and monetary policy. *J. Financ. Stab.* **6**(3), 121–129 (2010). <https://doi.org/10.1016/j.jfs.2009.07.001>
- Amidu, M., Wolfe, S.: The effect of banking market structure on the lending channel: evidence from emerging markets. *Rev. Financ. Econ.* **22**(4), 146–157 (2013). <https://doi.org/10.1016/j.rfe.2013.05.002>
- Aysun, U., Hepp, R.: Securitization and the balance sheet channel of monetary transmission. *J. Bank. Finance* **35**(8), 2111–2122 (2011)
- Bashir, U., Yugang, Y., Hussain, M.: Role of bank heterogeneity and market structure in transmitting monetary policy via bank lending channel: empirical evidence from Chinese banking sector. *Post-Communist Econ.* **32**(8), 1038–1061 (2020). <https://doi.org/10.1080/14631377.2019.1705082>
- Bean, C.R., Larsen, J.D., Nikolov, K.: Financial frictions and the monetary transmission mechanism: theory, evidence and policy implications. *Evid. Policy Implications* (2002)
- Beck, T., Levine, R., Loayza, N.: Finance and the sources of growth. *J. Financ. Econ.* **58**(1–2), 261–300 (2000). [https://doi.org/10.1016/S0304-405X\(00\)00072-6](https://doi.org/10.1016/S0304-405X(00)00072-6)

- Bernanke, B.S., Blinder, A.S.: Credit, money, and aggregate demand. In: National Bureau of Economic Research Cambridge, Mass., USA (1988)
- Boivin, J., Kiley, M.T., Mishkin, F.S.: How has the monetary transmission mechanism evolved over time? In *Handbook of Monetary Economics*, vol. 3, pp. 369–422. Elsevier (2010)
- Brissimis, S.N., Delis, M.D.: Identification of a loan supply function: a cross-country test for the existence of a bank lending channel. *J. Int. Financ. Mark. Inst. Money* **19**(2), 321–335 (2009). <https://doi.org/10.1016/j.intfin.2008.01.004>
- Cantero-Saiz, M., Sanfilippo-Azofra, S., Torre-Olmo, B., López-Gutiérrez, C.: Sovereign risk and the bank lending channel in Europe. *J. Int. Money Finance* **47**, 1–20 (2014). <https://doi.org/10.1016/j.jimonfin.2014.04.008>
- Dang, V.D., Nguyen, K.Q.B.: Monetary policy, bank leverage and liquidity. *Int. J. Manag. Finance* **17**, 619–639 (2020)
- Disyatat, P.: The bank lending channel revisited. *J. Money Credit Bank.* **43**(4), 711–734 (2011). <https://doi.org/10.1111/j.1538-4616.2011.00394.x>
- Ferreira, C.: The credit channel transmission of monetary policy in the European Union: a panel data approach. *Banks Bank Syst.* **5**, 230–240 (2010)
- Freedman, P.L., Click, R.W.: Banks that don't lend? Unlocking credit to spur growth in developing countries. *Dev. Policy Rev.* **24**(3), 279–302 (2006)
- Friedman, M., Schwartz, A.J.: Money and business cycles. In: *The State of Monetary Economics*, pp. 32–78. NBER (1965)
- Gambacorta, L.: Inside the bank lending channel. *Eur. Econ. Rev.* **49**(7), 1737–1759 (2005). <https://doi.org/10.1016/j.eurocorev.2004.05.004>
- Gambacorta, L., Marques-Ibanez, D.: The bank lending channel: lessons from the crisis. *Econ. Policy* **26**(66), 135–182 (2011). <https://doi.org/10.1111/j.1468-0327.2011.00261.x>
- Gertler, M.: *Finance, Growth, and Public Policy*, vol. 814. World Bank Publications, Washington (1991)
- Gibson, M.S.: The bank lending channel of monetary policy transmission: evidence from a model of bank behavior that incorporates long-term customer relationships (1997)
- Halsey, L.G.: The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* **15**(5), 20190174 (2019)
- Held, L., Ott, M.: On p-values and Bayes factors. *Annu. Rev. Stat. Appl.* **5**(1), 393–419 (2018)
- Hou, X., Wang, Q.: Implications of banking marketization for the lending channel of monetary policy transmission: evidence from China. *J. Macroecon.* **38**, 442–451 (2013). <https://doi.org/10.1016/j.jmacro.2013.07.004>
- Isakova, A.: Monetary policy efficiency in the economies of Central Asia. *Czech J. Econ. Finance (Finance a uver)* **58**(11–12), 525–553 (2008)
- Iturriaga, F.J.L.: More on the credit channel of monetary policy transmission: an international comparison. *Appl. Financ. Econ.* **10**(4), 423–434 (2000)
- Jimborean, R.: The role of banks in the monetary policy transmission in the new EU member states. *Econ. Syst.* **33**(4), 360–375 (2009). <https://doi.org/10.1016/j.ecosys.2009.08.001>
- Kashyap, A.K., Stein, J.C.: The impact of monetary policy on bank balance sheets. In: *Carnegie-rochester conference series on public policy* (1995)
- Kishan, R.P., Opiela, T.P.: Bank size, bank capital, and the bank lending channel. *J. Money Credit Bank.* **32**, 121–141 (2000)
- Kishan, R.P., Opiela, T.P.: Bank capital and loan asymmetry in the transmission of monetary policy. *J. Bank. Finance* **30**(1), 259–285 (2006). <https://doi.org/10.1016/j.jbankfin.2005.05.002>
- Leroy, A.: Competition and the bank lending channel in Eurozone. *J. Int. Financ. Mark. Inst. Money* **31**, 296–314 (2014). <https://doi.org/10.2139/ssrn.2341669>
- Lerskullawat, A.: Effects of banking sector and capital market development on the bank lending channel of monetary policy: an ASEAN country case study. *Kasetsart J. Soc. Sci.* **38**(1), 9–17 (2017). <https://doi.org/10.1016/j.kjss.2016.10.001>

- Maddaloni, A., Peydró, J.-L.: Bank risk-taking, securitization, supervision, and low interest rates: evidence from the Euro-area and the US lending standards. *Rev. Financ. Stud.* **24**(6), 2121–2165 (2011)
- Mishkin, F.S.: The channels of monetary transmission: lessons for monetary policy. In: National Bureau of Economic Research Cambridge, Mass., USA (1996)
- Mishra, P., Montiel, P., Sengupta, R.: Monetary transmission in developing countries: evidence from India. In: Ghate, C., Kletzer, K.M. (eds.) *Monetary policy in India*, pp. 59–110. Springer, New Delhi (2016). [https://doi.org/10.1007/978-81-322-2840-0\\_3](https://doi.org/10.1007/978-81-322-2840-0_3)
- Olivero, M.P., Li, Y., Jeon, B.N.: Banking competition and the lending channel: evidence from bank-level data in Asia and Latin America. In: 22nd Australasian Finance and Banking Conference (2009)
- Olivero, M.P., Li, Y., Jeon, B.N.: Competition in banking and the lending channel: evidence from bank-level data in Asia and Latin America. *J. Bank. Finance* **35**(3), 560–571 (2011). <https://doi.org/10.1016/j.jbankfin.2010.08.004>
- Opiela, T.P.: Differential deposit guarantees and the effect of monetary policy on bank lending. *Econ. Inquiry* **46**(4), 610–623 (2008). <https://doi.org/10.1111/j.1465-7295.2007.00100.x>
- Peek, J., Rosengren, E.S.: *The Role of Banks in the Transmission of Monetary Policy*. Oxford University Press, Oxford (2010)
- Perera, A., Ralston, D., Wickramanayake, J.: Impact of off-balance sheet banking on the bank lending channel of monetary transmission: evidence from South Asia. *J. Int. Financ. Mark. Inst Money* **29**, 195–216 (2014). <https://doi.org/10.1016/j.intfin.2013.12.008>
- Pham, K.D.: Financial crisis and diversification strategies: the impact on bank risk, and performance (2021). 739898418
- Rybczynski, T. M.: A new look at the evolution of the financial system. In: Revell, Jack (ed.) *The Recent Evolution of Financial Systems*, pp. 3–15. Palgrave Macmillan UK, London (1997). [https://doi.org/10.1007/978-1-349-14192-0\\_1](https://doi.org/10.1007/978-1-349-14192-0_1)
- Sáiz, M.C., Azofra, S.S., Olmo, B.T., Gutiérrez, C.L.: A new approach to the analysis of monetary policy transmission through bank capital. *Financ. Res. Lett.* **24**, 95–104 (2018)
- Sanfilippo-Azofra, S., Torre-Olmo, B., Cantero-Saiz, M., López-Gutiérrez, C.: Financial development and the bank lending channel in developing countries. *J. Macroecon.* **55**, 215–234 (2018). <https://doi.org/10.1016/j.jmacro.2017.10.009>
- Vo, X.V., Nguyen, P.C.: Monetary policy and bank credit risk in Vietnam pre and post global financial crisis. In: *Risk Management Post Financial Crisis: A Period of Monetary Easing*. Emerald Group Publishing Limited (2014). <https://doi.org/10.1108/S1569-375920140000096011>
- Yang, J., Shao, H.: Impact of bank competition on the bank lending channel of monetary transmission: evidence from China. *Int. Rev. Econ. Finance* **43**, 468–481 (2016). <https://doi.org/10.1016/j.iref.2015.12.008>
- Zhan, S., Tang, Y., Li, S., Yao, Y., Zhan, M.: How does the money market development impact the bank lending channel of emerging countries? A case from China. *North Am. J. Econ. Finance* **57**, 101381 (2021). <https://doi.org/10.1016/j.najef.2021.101381>



# A Bayesian Binary Logistic Regression Approach to Identifying Factors Affecting the Households' Use Level of Financial Products/Services in Vietnam

Huong Thi Thanh Tran<sup>(✉)</sup>

Faculty of Accounting and Auditing, Banking Academy, Hanoi, Vietnam  
huongttt76@hvn.h.edu.vn

**Abstract.** This study uses the Bayesian Binary Logistic Regression method with data obtained from the Vietnam Household Living Standards Survey (VHLSS) conducted by the General Statistics Office of Vietnam (GSO) in 2018 to identify the factors affecting the households' use level of financial products/services (FS) in Vietnam. The research results show that households living in urban areas are more likely to have bank accounts and use ATM cards than those living in rural areas. Household size increases the household's probability of having bank accounts and ATM cards. Households with access to information and communications have a higher probability of having bank accounts and using ATM cards than those without access. The education level of the householder increases the household's probability of having bank accounts and using ATM cards. The employed householders have a higher probability of having bank accounts and using ATM cards than the unemployed ones. Male householders are less likely to have bank accounts and use credit cards than female householders. The study also provides some recommendations to promote the households' use level of FS in Vietnam.

**Keywords:** Financial products/services · Households · Bayesian binary logistic regression

## 1 Introduction

Financial inclusion (FI) is defined as the process of supply of FS in a reasonable, convenient and timely manner to all members of society, especially the poor and the vulnerable (Mohan 2006; Rangarjan 2008; Ajide 2015). FI has opened up opportunities for people to have access to FS. The increase in the number of households and businesses using banking services has brought positive contributions to the economic growth and poverty reduction of countries. Finding solutions that can break down the barriers faced by the poor to access formal FS has become a great concern of researchers, governments, and financial institutions. Recognizing the importance of FI to economic growth and development, Vietnam has also issued a lot of policies to promote FI such as the Government's Decree No. 55/2015/ND-CP dated June 9, 2015 on Credit policy for agricultural and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 612–626, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_43](https://doi.org/10.1007/978-3-031-35763-3_43)

rural development; the Prime Minister's Decision No. 2195/QĐ-TTg dated December 6, 2011 on promulgating the Scheme on building and development of a microfinance system in Vietnam up to 2020. The Prime Minister's Decision No. 986/QĐ-TTg dated August 8, 2018 on the Strategy for development of Vietnam's banking sector by 2025, with vision to 2030, defining more clearly the goal of achieving FI by 2030, ensuring that all people and businesses have an opportunity to fully and conveniently access high-quality banking and FS, making positive contributions to sustainable development.

There have been many studies on FI in terms of measuring or assessing the impacts of FI on economic growth, poverty and inequality reduction (Clamara et al. 2014; Tran and Le 2021). Several studies have explored the factors affecting the development of FI in some countries and regions, typically some studies by Allen et al. (2012), Ajjide (2017), Kaur (2017), Lenka and Barik (2018). However, there are relatively few studies evaluating the factors which have effect on FI at a micro (household) perspective. Moreover, in Vietnam, to the best of the author's knowledge, there are no studies that explore the factors affecting FI at the household level. This research paper fills the research gap by identifying the factors affecting the households' use level of FS, thereby making recommendations to increase the level of households' access to FS in Vietnam.

To identify the factors affecting the households' use level of FS, the author uses the Bayesian Binary Logistic Regression method with data obtained from VHLSS 2018 conducted by GSO. The research results show that households living in urban areas are more likely to have bank accounts and use ATM cards than those living in rural areas. Household size increases the household's probability of having bank accounts and using ATM cards. Households with access to information and communications have a probability of having bank accounts and using ATM cards higher than those without access. The education level of householder increases the household's probability of having bank accounts and using ATM cards. The employed householders have a probability of having bank accounts and using ATM cards higher than the unemployed ones. Male householders are less likely to have bank accounts and use credit cards than female householders.

The structure of the research paper, in addition to the introduction and conclusion, includes literature review, research models and data sources, estimation results and discussion.

## 2 Literature Review

Studies on factors affecting FI approach this problem from both macro (national, regional) and micro (household) perspectives. From a macro perspective, Allen et al. (2012), Fungáčová and Weill (2014), and Demirgüç-Kunt et al. (2013) all find a link between demand-side factors (individual characteristics) and FI. Specifically, on a global scale, Allen et al. (2012) find that the percentage of the educated, elderly, employed, married or separated people having accounts and savings at a formal financial institution is higher than that of other group. The research by Fungáčová and Weill (2014) in China also indicates that the poor, the less educated and the younger have less access to FS. The research by Demirgüç-Kunt et al. (2013) with samples of 98 developing and emerging countries shows that there is a gender gap between those with accounts,

savings and loans. On the supply side, the papers by Chakravarty and Pal (2010) and Kuma (2013) represent that the system of transaction offices (the number of branches and employees) is a decisive factor for FI while penetration of branch network, and credit are two important strategies to promote FI. Sousa (2015), using panel data of 90 developing and emerging countries from the Global Findex database 2011, concludes that the explanatory factors for FI include macroeconomic variables such as GDP per capita, inflation, and supply-side variables such as real interest rates, bank credit to the private sector, Z-score, and financial market penetration. Similarly, Ajide (2017), using the SGMM (SYS-GMM) model with array data of 18 sub-Saharan African countries, emphasizes the importance of macro variables such as GDP per capita, inflation, institutional quality, and supply-side variables such as bank concentration and Z-score in promoting FI. Singh and Kodan (2011) examine the impacts of factors such as education level, unemployment rate, male to female ratio (demand side), GDP per capita, and urbanization rate (macro variables) on FI. They conclude that macroeconomic factors are the main contributors to FI. Kaur (2017) also shows that urbanization rate, education level, and GDP all have effect on FI; however, the contribution of education level to FI is lower than other variables. The empirical research of Lenka and Barik (2018) on control variables shows that income and education level are positively related to FI while rural population size and unemployment rate are negatively related to FI. Moreover, Meskoub (2018) also argues that the cause of financial exclusion is due to low income, high unemployment rate, low education level, poor living conditions mainly in small towns and countryside.

From a micro perspective, Devlin (2005) focuses on the determinants of the use of banking services. Based on data collected from face-to-face interviews with 15880 persons in the UK in 2000, the study finds that the factors contributing to FI include employment status, household income, and housing tenure. Also using micro data to examine demand-side factors, Clamara et al. (2014) use probit models to explore individual characteristics that affect the choice to use formal FS. Their research indicates that the vulnerable groups, including women, young people, and those living in rural areas, often have difficulty accessing the banking and FS. For small businesses, the business type as well as the education level of the head are the factors affecting access to FS. Examining both supply- and demand-side factors affecting the households' use of FS based on micro data from 123 countries, Allen et al. (2012) use probit models and indicate that the degree of FI is negatively correlated with demand-side factors such as low income, and living in rural areas while a positive correlation is found between FI and households' perception of supply-side factors such as lower costs of banking services, proximity to bank branches and transaction offices, and reduced documentation requirements.



### 3 Research Models and Data Sources

#### 3.1 Research Models

To identify the effects on the households' use level of FS, on the basis of the literature review of the factors' impact on FI and the data conditions possible to be collected, the authors propose the following binary logistic regression model:

$$\log\left(\frac{P_{it}}{1 - P_{it}}\right) = \beta_0 + \beta_1 AREA_{it} + \beta_2 HSIZE_{it} + \beta_3 IAP_{it} + \beta_4 SEX_{it} + \beta_5 EDU_{it} + \beta_6 JOB_{it} + u_{it} \quad (1)$$

where:  $P_i$  is the probability that the variable  $FI$  (the households' use level of financial products/services) takes the value "1" given the values of input variables:  $AREA$ ,  $HSIZE$ ,  $IAP$ ,  $SEX$ ,  $EDU$ ,  $JOB$ .

$$P_i = P(FI_i = 1|Z_i) = \frac{1}{1 + e^{-Z_i}}$$

with:

$$Z_{it} = \beta_0 + \beta_1 AREA_{it} + \beta_2 HSIZE_{it} + \beta_3 IAP_{it} + \beta_4 SEX_{it} + \beta_5 EDU_{it} + \beta_6 JOB_{it} \quad (2)$$

$1 - P_i$  is the probability that the variable  $FI$  takes the value "0";  $(P_i/1 - P_i)$  is the (odds) ratio of the households' probability of using FS ( $FI = 1$ ) to that of unuse of FS. Because  $FI$  reflects the households' use level of FS, we propose two variables FI: Household with bank accounts ( $BAC$ ), Household with ATM cards ( $ATM$ ); The independent variables used in the model include 6 variables reflecting the characteristics of household or householder: area of residence of household ( $AREA$ ), household size ( $HSIZE$ ); household's access to information ( $IAP$ ); sex of householder ( $SEX$ ), education level of householder ( $EDU$ ); job of householder ( $JOB$ );  $u_{it}$  and  $\varepsilon_{it}$ : random errors;  $i$  is the surveyed household,  $t$  is year.

According to Anjullo and Haile (2018), the analysis of logistic regression model is based on estimating the model's parameters through Maximum Likelihood Estimation and calculating by the expectation maximization algorithm. However, in the bayesian approach, the inference of the model parameters is conducted on the basis of their posterior distribution which is the combination of the likelihood function of observed data and information from previous studies or personal experiences that are known as prior distribution (Howson and Urbach 2006). Therefore, this study uses the Bayesian binary logistic regression model to estimate the factors' influence on the households' use level of financial products/services. Compared with the frequency approach, the Bayesian approach is a robust estimator by combining priori information about model parameters with observed data to form a posterior model of interest while the frequency analysis relies entirely on data. Furthermore, according to Nguyen Ngoc Thach (2020), the confidence interval in the frequency approach has no clear probabilistic explanation compared with the posterior confidence interval in the Bayesian analysis. The likelihood

function on a sample dataset with the size of  $n$  subjects, the likelihood function for data  $FI = (FI_1, FI_2, \dots, FI_n)^T$  is:

$$\begin{aligned}
 prob(FI|\beta) &= L(\beta|FI) = \prod_{i=1}^n [P_i^{FI_i} (1 - P_i)^{(1-FI_i)}] \\
 &= \prod_{i=1}^n \left( \frac{e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}}{1 + e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}} \right)^{FI_i} \\
 &\quad \left( 1 - \frac{e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}}{1 + e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}} \right)^{(1-FI_i)}
 \end{aligned} \tag{3}$$

Furthermore, one of the prerequisites in any Bayesian analysis is the prior distribution for logistic regression parameters is normal distribution with mean  $\mu_j$  and with variance  $\sigma_j^2$  and has the form:  $\beta_j \sim N(\mu_j, \sigma_j^2)$ . The prior distribution for logistic regression parameters has the form (4):

$$P(\beta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{1}{2} \left( \frac{\beta_j - \mu_j}{\sigma_j^2} \right)^2 \right\} \tag{4}$$

The most common choice for the prior mean  $\mu_j$  is 0 for all coefficients and large enough prior variance  $\sigma_j^2$ . Therefore, in this study, due to the lack of past information for the prior distribution of regression coefficients, the non-informative normal prior distribution with the prior distribution parameters of mean 0 and variance 1000 was considered. Then, for the choice of non-informative independent normal priors and likelihood function for the data, posterior distribution of model parameters which is the product of the Eqs. (4) and (3) has the form (5 and 6):

$$P(\beta|FI) = \prod_{j=0}^p [P(\beta_j)] * \prod_{i=1}^n [L(\beta|FI)] \tag{5}$$

$$\begin{aligned}
 &= \prod_{j=0}^p \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{1}{2} \left( \frac{\beta_j - \mu_j}{\sigma_j^2} \right)^2 \right\} \\
 &* \prod_{i=1}^n \left[ \frac{\left( \frac{e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}}{1 + e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}} \right)^{FI_i}}{\left( 1 - \frac{e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}}{1 + e^{\beta_0 + \beta_1 AREA + \beta_2 HSIZE + \beta_3 IAP + \beta_4 SEX + \beta_5 EDU + \beta_6 JOB}} \right)^{(1-FI_i)}} \right]
 \end{aligned} \tag{6}$$

The posterior distribution above is a complex function of the parameters, and numerical methods are needed to obtain the marginal posterior distribution for each model parameter. The most common method of simulating from a general posterior distribution is to use the Markov Chain Monte Carlo (MCMC) simulation method (Spiegelhalter et al. 1996). Among the MCMC methods, the most commonly used method is the Metropolis-Hastings method and the Gibbs sampling method. Gibbs sampling method

is considered as a special case, with special properties of Metropolis-Hastings algorithm (Gelfand et al. 1990), which is a highly efficient method despite requiring relatively large computational resources. In this research, the author uses the Bayesian approach through Random-Walk Metropolis – Hastings algorithm and Gibbs sampling method to find out the factors' influence on the households' use level of FS.

Measurement of variables used in the model is presented in the Table 1 below:

**Table 1.** Measurement of variables

Dimensions	Name of variables	Abbreviations	Measurement
Use level of FS(FI)	Household has bank accounts	BAC	=1 if household has bank accounts =0 if household does not have bank accounts
	Household uses ATM cards	ATM	=1 if household uses ATM cards =0 if household does not use ATM cards
Area	Area of residence of household	AREA	=1 if living in urban area =0 if living in rural area
Size	Household size	HSIZE	Total number of members in the household
Access to information	Household's use of telecommunications services and assets for access to information	IAP	=1 if the household has a member using telephone and internet subscriptions, and the household has one of the assets for access to information (television, radio, computer) =0 if otherwise
Sex	Sex of householder	SEX	=1 if householder is female =0 if householder is male
Education	Education level of householder	EDU	Number of years of schooling of the householder
Job	Job of householder	JOB	=1 if the householder has a job =0 if the householder has no job

### 3.2 Data Sources

The data used in this study are collected from the results of VHLSS conducted by GSO in 2018. The VHLSS 2018 is a sampling survey, including 72054 households

(37596 households collected information on income, and consumer price index weight; 9396 households collected information on income and expenditure; 25059 households collected information on consumer price index) selected from 4177 locations of the master sample. The sampling frame is selected from the 15% sampling frame of the 2009 Vietnam Population and Housing Census updated as the survey is conducted. The data used in this study belong to the sample of 9396 households that are fully surveyed with information about their living standards. Although the survey sample included 9396 households, some households did not answer some questions. Therefore, to obtain the best estimate, the study excludes the missing observations (no answers); hence, the sample size used in this estimate includes 8166 households.

## 4 Estimation Results and Discussion

Descriptive statistics of variables are presented in Table 2. According to the data in Table 2, out of 8166 surveyed households, the percentage of households with bank accounts and ATM cards is quite low, 30.2% and 39.8%, respectively. The percentage of households living in rural areas is twice as high as those living in urban areas. On average, a household has 3.71 members. Most of the households in the survey sample have means of information access (accounting for 95.1%). The percentage of households with male householder is three times as high as that with female householder. The average number of years of schooling of householder is quite low (7.94 years, nearly graduated from secondary school). The percentage of employed householders is quite high, accounting for 85.4%.

Two important criteria to evaluate the efficiency of the MCMC sampling algorithm in Bayesian models are the acceptance rate and the MCMC sample efficiency. The calculation results of this study show that the models corresponding to the dependent variables reflecting the households' use level of FS (*BAC*, *ATM*) have the acceptance rate greater than 0.15 (0.20, 0.27 respectively). Meanwhile, according to Roberts and Rosenthal (2001), the acceptance rate in the range of 0.15 - 0.5 is optimal. Therefore, the MCMC sample of the regression model has reached the allowable acceptance rate. The minimum, average and maximum efficiencies of the models corresponding to the dependent variables (*BAC* and *ATM*) are all greater than the alarm level, 0.01, (The minimum, average and maximum efficiencies of *BAC* are 0.013, 0.026, 0.041, respectively; the ones of *ATM* are 0.013, 0.028, 0.052, respectively). Furthermore, according to Hosmer et al. (2013), the values of the Monte Carlo Standard Error (MCSE) should be less than 5%. In our study, the MCSE values of the posterior mean in all models are less than 5% of its posterior standard error (Table 3). This means that convergence and accuracy of the posterior estimates for the regression parameters are achieved. In order to ensure that the Bayesian inference based on the MCMC sample is reasonable, the author continues to test the MCMC convergence of the parameter estimates by the visual graphical diagnosis method. Figures 1 and 2 in the appendix (corresponding to the dependent variables *BAC*, *ATM*) show that the trace plots do not have trends, fluctuate around the mean values; therefore, the MCMC chain is stationary, that is, the convergence conditions are met; the autocorrelation plots fall off fast ( $<40$ ); the histograms and the density plots show that simulation of the normal distribution shape of the parameters, the shape of the histogram

**Table 2.** Descriptive statistics

	Total sample
Number of households	8166
<b><i>BAC</i></b>	
1: Household has bank accounts	2469 (30.2%)
0: Household does not have bank accounts	5697 (69.8%)
<b><i>ATM</i></b>	
1: Household uses ATM cards	3252 (39.8%)
0: Household does not use ATM cards	4914 (60.2%)
<b><i>AREA</i></b>	
1: Urban area	2548 (31.2%)
0: Rural area	5618 (68.8%)
<b><i>HSIZE</i></b>	
Mean (SD)	3.71 (1.55)
Median [Min, Max]	4.00 [1.00, 15.0]
<b><i>IAP</i></b>	
1: Household has means of information access	7764 (95.1%)
0: Household does not have means of information access	402 (4.9%)
<b><i>SEX</i></b>	
1: Male	6157 (75.4%)
0: Female	2009 (24.6%)
<b><i>EDU</i></b>	
Mean (SD)	7.94 (4.8)
Median [Min, Max]	9.0 [0, 23]
<b><i>JOB</i></b>	
1: Householder has a job	6974 (85.4%)
0: Householder has no job	1192 (14.6%)

is uniform, so it can be concluded that the Bayesian inference is stable. Thus, through the MCMC convergence test results, it can be concluded that the MCMC chain satisfies the convergence conditions.

Table 3 presents the estimates (posterior mean, Odds ratio, Monte Carlo error, posterior median, and 95% Credible interval of posterior mean) of the factors affecting the households' use level of FS obtained from the Bayesian binary logistic regression model. Based on the Bayesian 95% confidence interval, the predictors including area of residence of household, household size, household's access to information, sex, education level and job of householder, are found to be statistically significant factors affecting the households' use level of FS.

**Table 3.** Simulation results of the model

Predictors	Posterior mean	Odds ratio	Posterior standard error	MCSE	Posterior median	95% Credible interval of posterior mean	
<b><i>BAC</i></b>							
AREA	1.0692	2.9180	0.0593	0.0043	1.0694	0.9576	1.1793
HSIZE	0.2354	1.2656	0.1757	0.0013	0.2358	0.2005	0.2694
IAP	0.1580	1.1813	0.1315	0.0089	0.1535	-0.1063	0.4163
SEX	-0.3495	0.7067	0.0681	0.0036	-0.3474	-0.4863	-0.2219
EDU	0.1354	1.1450	0.0062	0.0003	0.1354	0.1230	0.1480
JOB	0.1287	1.0162	0.0811	0.0070	0.0142	-0.1533	0.1677
Cons	-3.1733	0.0424	0.1613	0.0089	-3.1700	-3.4920	-2.8446
<b><i>ATM</i></b>							
AREA	0.9111	2.4909	0.0562	0.0025	0.9119	0.7965	1.0207
HSIZE	0.3469	1.4149	0.0177	0.0011	0.3474	0.3131	0.3803
IAP	0.0686	1.0782	0.1154	0.0066	0.0672	-0.1601	0.2949
SEX	-0.4862	0.6161	0.0614	0.0052	-0.4829	-0.6085	-0.3679
EDU	0.1532	1.1656	0.0060	0.0004	0.1531	0.1416	0.1651
JOB	0.1271	1.1385	0.0727	0.0046	0.1287	-0.0112	0.2654
Cons	-3.0938	0.0459	0.1567	0.0096	-3.0925	-3.3946	-2.7863

The estimation results of Table 3 show that households living in urban areas have a probability of using FS higher than those living in rural areas. Households living in urban areas are 2.918 times more likely to have bank accounts than those living in rural areas (odds: 2.918, 95% confidence interval: 0.9576; 1.1793). Households living in urban areas are 2.4909 times more likely to use ATM cards than those living in rural areas (odds: 2.4909, confidence interval: 0.7965, 1.0207). The research results are also similar to the studies by Allen et al. (2012), Clamara et al. (2014), which indicate that households living in rural areas have lower level of use of FS. The percentage of adults with bank accounts is an important indicator of FI. According to calculations based on supply-side data by the State Bank of Vietnam (SBV), this percentage in Vietnam is currently 64%, lower than neighboring countries such as China, Malaysia, Thailand, which have now reached more than 80%. The number of people without bank accounts is mainly concentrated in rural, remote and isolated areas. According to World Bank Group (2017), the percentage of people in rural areas with accounts in Vietnam is much lower than the world's average data. In Vietnam, 25.2% of people in rural areas have bank accounts while on average this rate reaches 66% in the world. According to SBV's demand side survey data in May 2019, 34% of adults living in rural areas own bank accounts, much lower than those living in urban areas (57%); 90% of urban people take nearly 15 min to get to the nearest financial service providers while this rate is much lower in rural areas,

less than 40% (SBV, 2021). Although Vietnam's ATM system has developed rapidly, its distribution is uneven. The number of ATMs is concentrated mainly in urban areas where payment services are developed while there are only a few ATMs in rural and mountainous areas. According to the data of World Bank Group (2017), compared with some countries in the region, Vietnam's coverage of financial access points is still quite low; in 2017, the number of ATMs per 100,000 adults in Vietnam was 24.34; the number of bank branches per 100,000 adults was 3.4 (while the rates were respectively 117.28 and 11.88 in Thailand; 46.75 and 10.05 in Malaysia; 55.61 and 116.89 in Indonesia). Thus, in order to increase the use level of financial products/services, Vietnam needs to pay attention to policies to promote financial development in rural areas.

The data of Table 3 show that household size increases the households' probability of using FS. When the number of members is more than 1 person, the probability of having bank accounts increases to 1.2656 times (odds: 1.2656, confidence interval: 0.2005, 0.2694) and the probability of using ATM cards increases to 1.4149 times (odds: 1.4149, confidence interval: 0.3131, 0.3803). Households with means of access to information and communications have a probability of having bank accounts and using ATM cards 1.1813 and 1.0782 times, respectively, as high as those without access. Male householders have a probability of having bank accounts 29.33% as low as female householders (odds: 0.7067, confidence interval is:  $-0.4863$ ,  $-0.2219$ ). Male householders have a probability of using ATM cards 19.39% as low as female householders (odds: 0.6161, confidence interval:  $-0.6085$ ,  $-0.3679$ ). The education level of householder increases the household's probability of having bank accounts and using ATM cards. When the number of years of schooling of householder is higher than 1 year, the probability of having bank accounts increases to 1.1450 times (odds: 1.1450, confidence interval: 0.1230, 0.1480) and the probability of using ATM cards increases to 1.1656 times (odds: 1.1656, confidence interval: 0.1416, 0.1651). These research results are also similar to those by Fungáčová and Weill (2014), Lenka and Barik (2018) which show that education level has a positive impact on the level of use of financial products/services. According to World Bank Group (2017) there is a quite large difference in the rate of people with accounts in the group with high school degree compared to the group with primary school degree or below. This rate is 78.7% and 55.9% in the world, 42.4% and 13.0% in Vietnam. Figures on the percentage of account holders by education level indicate that people with lower levels of education tend to have lower access to formal financial products. According to SBV (2021), the account ownership rate among non-schoolers is very low, only 12% in 2019. According to the data of Table 3, the employed householders have a probability of having bank accounts 1.0162 times as high as the unemployed ones, and the probability of using ATM cards 1.1385 times as high as the unemployed ones. This finding is also similar to the research results of Davlin (2005), Allen et al. (2012) which show that employment status affects the level of use of financial products/services. In order to develop FI, in the future, Vietnam needs to focus on solutions to improve education levels and create job opportunities for people.

## 5 Conclusion

This study uses Bayesian binary logistic regression model to indentify the factors affecting the households' use level of FS. The research results show that households living in urban areas are more likely to have bank accounts and use ATM cards than those living in rural areas. Household size increases the household's probability of having bank accounts and using ATM cards. Households with access to information and communications have a probability of having bank accounts and using ATM cards higher than those without access. The education level of householder increases the household's probability of having bank accounts and using ATM cards. Male householders are less likely to have bank accounts and use credit cards than female householders. The employed householders have a probability of having bank accounts and using ATM cards higher than the unemployed ones. Thus, in order to promote the use of FS by households, in the coming time, Vietnam needs to focus on the following solutions:

Firstly, Vietnam needs to encourage banks and financial institutions to expand the coverage of branches, transaction offices, and ATM systems to rural and remote areas. According to the Findex survey 2017, many respondents also say that the main reason why they do not have an account is that the access to FS is still not convenient. In addition to traditional business models, Vietnam needs to promote new business models based on digital technology such as mobile banking, agent banking; as well as products and services such as e-wallets, mobile money, to create new transaction methods, helping people, especially those living in rural, remote and isolated areas easy to access FS. Besides, the promotion of application of digital technology and innovation in the banking and finance sector such as QR code payment, electronic customer identification (e-KYC), open data sharing via API (Open API), etc., will help remove barriers of geographical distance, and people will have more favorable conditions to access and use FS.

Secondly, the government needs to promote information and communications affairs to households. Step up education programs, and disseminate financial knowledge to all people and enterprises. Organize training courses in professional skills for staff in charge of information and communications works at the grassroots; priority is given to staff at commune and village levels. Support audiovisual media for poor households living on offshore islands; poor households belonging to ethnic minorities or in extremely difficult communes; continue to maintain support policies to improve the education level of people.



Appendix

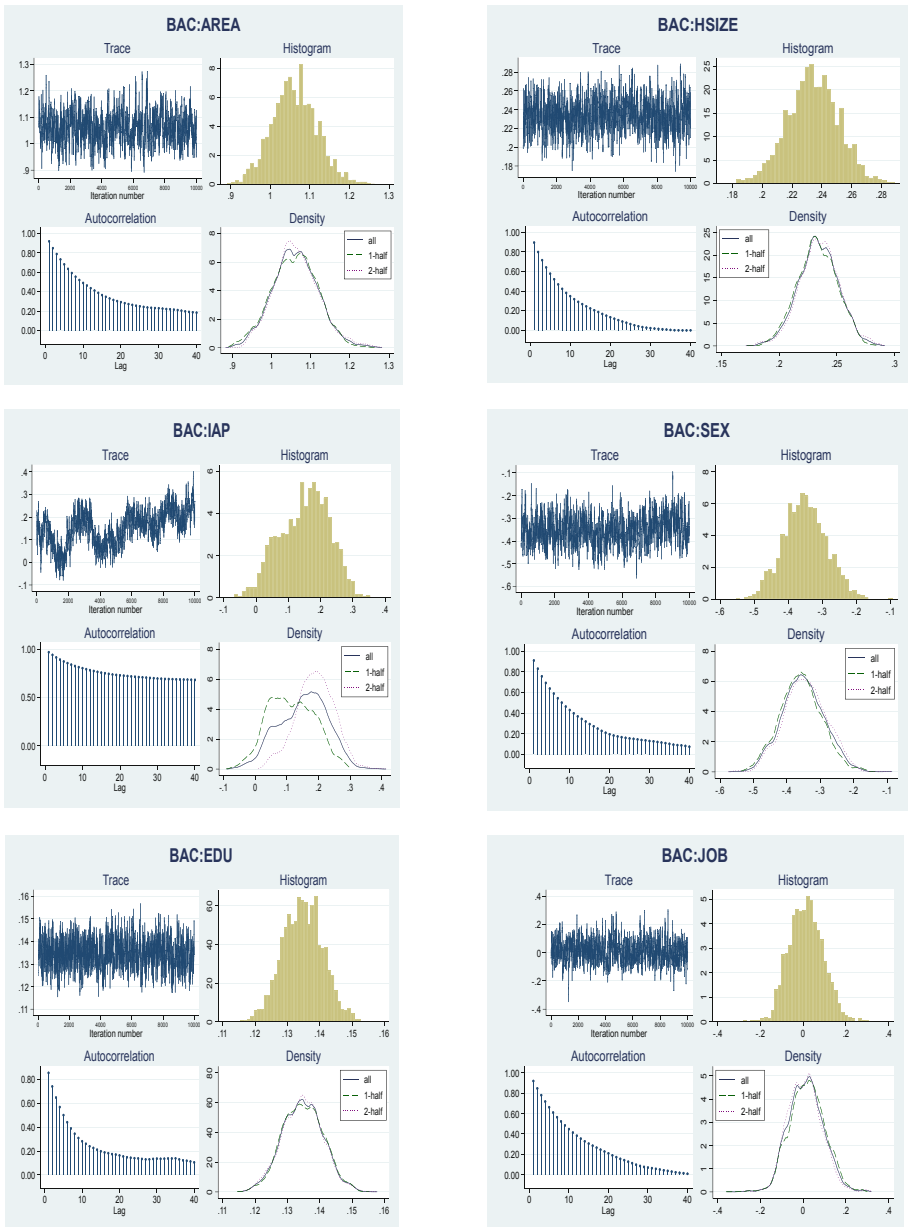


Fig. 1. Convergence diagnostics of MCMC chains for BAC variable

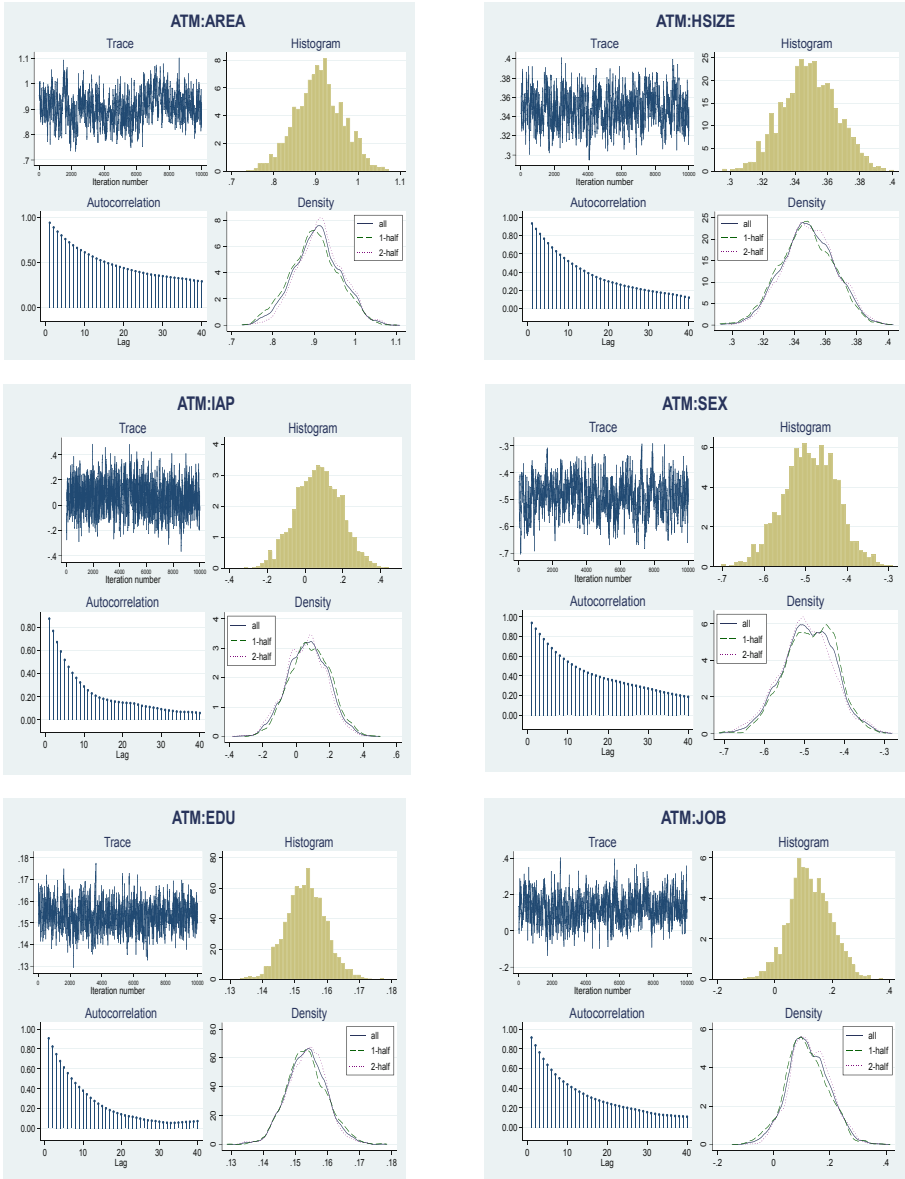


Fig. 2. Convergence diagnostics of MCMC chains for ATM variable

### References

Ajide, F.: Financial inclusion and Rural Poverty Reduction: Evidence from Nigeria. *Int. J. Manage. Sci. Hum.* **3**(2), 190–203 (2015)  
Ajide, K.B.: Determinants of financial inclusion in Sub-Saharan Africa countries: does institutional infrastructure matter? *CBN J. Appl. Stat.* **8**(2), 69–89 (2017)

- Allen, F., Demiguc-Kunt, A.I., Klapper, L., Peria, M.S.M.: The foundations of financial inclusion. Understanding Ownership and Use of Formal Accounts, Policy Research working paper, No. WPS 6290. Washington, DC: World Bank Group (2012)
- Anjullo, B.B., Haile, T.T.: A Bayesian binary logistic regression approach in identifying factors associated with exclusive breastfeeding practices at Arba Minch Town, South Ethiopia. *Adv. Res.* **17**(5), 1–14 (2018)
- Chakravarty, S.R., Pal, R.: Measuring financial inclusion: an axiomatic approach. Working paper, Mumbai: Indira Gandhi Institute of Development Research, No. Wp-2010–003 (2010)
- Clamara, N., Pena, X., Tuesta, D.: Factors that matter for financial inclusion: Evidence from Peru, BBVA Research Working Paper, No.14/09 (2014)
- Demirguc-Kunt, A., Klapper, L., Singer, D.: Financial inclusion and legal discrimination against women: Evidence from Developing Countries, Policy Research Working Paper, No. WPS6416. Washington, DC: World Bank Group (2013)
- Delvin, J.F.: A detailed study of financial exclusion in the UK. *J. Consum. Policy* **28**, 75–108 (2005)
- Fungáčová, Z., Weill, L.: Understanding Financial inclusion in China, Bank of Finland Discussion Papers, No. 10.201 4 (2014)
- Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F.: Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Stat. Assoc.* **85**(412), 972–985 (1990)
- Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. Wiley, Hoboken (2013)
- Howson, C., Urbach, P.: Scientific Reasoning: The Bayesian Approach. Open Court Publishing, Chicago (2006)
- Kaur, J.: Factors affecting financial inclusion: a case study of Punjab. *Int. J. Adv. Res. Dev.* **2**(6), 422–426 (2017)
- Kumar, N.: Financial inclusion and its determinants: evidence from India. *J. Financ. Econ. Policy* **5**(1), 4–19 (2013)
- Lenka, S.K., Barik, R.: Has expansion of mobile phone and internet use spurred Financial inclusion in the SAARC countries? *Financ. Innov.* **4**(1), 1–19 (2018)
- Meskoub, M.: Financial services in the EU: Is there a problem of financial exclusion? Working Paper, No. 638 (2018)
- Mohan, R.: Economic growth, financial deepening and Financial inclusion”, Annual Bankers’ Conference 2006, Hyderabad, 3 November (2006)
- Nguyen Ngoc, T.: Elasticity of substitution between capital and labor: estimation and implications for output growth of Vietnam’s non-financial firms. *J. Int. Econ. Manage.* (128), 88-108 (2020)
- Rangarajan, C.: Report of the committee on financial inclusion, Ministry of Finance, Government of India (2008)
- Roberts, G.O., Rosenthal, J.S.: Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* **16**(4), 351–367 (2001)
- Singh, K., Kodan, A.S.: Financial inclusion, development and its determinants: an empirical evidence of Indian states. *Asian Econ. Rev. J. Indian Inst. Econ.* **53**(1), 115–134 (2011)
- Sousa, M.M.: Financial inclusion and Global regulatory standards: an empirical study across developing economies, Centre for International Governance Innovation Working Paper, No. 7 (2015)
- Spiegelhalter, D., et al.: BUGS 0.5: Bayesian inference using Gibbs sampling manual (version II). MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK, pp. 1–59 (1996)
- State Bank of Vietnam, Inquiries about Financial inclusion, Vietnamese Women’s Publishing House (2021)
- The Government’s Decree No. 55/2015/ND-CP dated June 9, 2015 on Credit policy for agricultural and rural development

The Prime Minister's Decision No. 2195/QĐ-TTg dated December 6, 2011 on promulgating the Scheme on building and development of a microfinance system in Vietnam up to 2020

The Prime Minister's Decision No. 986/QĐ-TTg dated August 8, 2018 on the Strategy for development of Vietnam's banking sector by 2025, with vision to 2030

Tran, H.T.T., Le, H.T.T.: The impact of financial inclusion on poverty reduction. *Asian J. Law Econ.* **12**(1), 95–119 (2021)

World Bank Group (2017). The Global Findex Database (2017). <https://www.worldbank.org/en/publication/globalfindex/Data>



# Impacts of Global Pandemics, Financial Crises, and Oil Price Shocks on Japanese Stock Market

Rongchai Tansuchat and Chaiwat Klinlampu<sup>(✉)</sup>

Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University,  
Chiang Mai 50200, Thailand  
chaiwatklinlampu@gmail.com

**Abstract.** Several financial and health crises pushed the global economy and stock market into a severe recession. Moreover, the global economic slowdown and contraction have decreased demand for crude oil, putting more pressure on stock markets. This study investigates the impacts of financial and health crisis events and oil prices on the Japanese stock market due to Japanese stock reacted to a large drop after the crisis occurred in many times. The autoregressive distributed lag (ARDL) model was adopted to find the result. Furthermore, by separating the changes in oil prices into positive and negative shocks, the short- and long-run asymmetries on the oil-stock nexus are also explored. Our findings show strong evidence of the negative impact of financial and health crisis events and oil price shocks in the short run. However, these short-run effects do not last into long-run effects. Besides, the result reveals the higher impact of a health crisis compared to a financial crisis.

**Keywords:** ARDL · Health crisis · Financial crisis · Japanese stock market · Oil price shocks

## 1 Formulation of the Problem

Many viral pandemics have occurred recently in the last two decades, such as Bird flu (2005), Swine flu (2009), Ebola (2014), and COVID-19. These outbreaks have contributed a large impacts and significant impact on the world in many aspects: economics, environment, and society [16]. However, none of these aspects rapidly respond to these outbreaks as the financial sector has. Although health crises have become a vital event generating impacts on the financial sector, it has been well known that the financial or economic crisis is another event that contributes a large negative impact on financial markets [22]. Moreover, the global economy's recession has reduced the demand for crude oil and increased pressure on the financial markets.

Financial markets, a part of the financial sector, are generally participated by important corporates for economic development such as banks, insurance companies, real estate developers, mortgage lenders, financial companies, investors,

speculators, etc. These entities and individuals are often one of the largest segments of the stock market, which is a type of financial market [17]. The stock market is widely perceived as an indicator of a country's economic performance. According to [14], the stock market is significant for economic development. As a result, consistent market performance is necessary to achieve long-term growth.

As financial markets are affected by health and financial crises and oil, stock market performance is also affected. There is much evidence confirming that these two types of crises brought higher unemployment and pushed the market into a severe recession ([1, 13, 22]). The emergence of health crises frequently restricts economic activity and reduces corporate profits [12], whereas financial crises cause volatility and raise the risk of financial assets [4].

In terms of oil, it is usually considered one of the biggest economic drivers [3]. Hence, it is reasonable to have oil as another factor leading to the performance of the stock market. There are three theoretical approaches explaining the impact of oil price on stock market returns. First is the stock valuation approach. This approach explained that crude oil could be viewed as a production factor. Therefore, the movement of the oil price can affect the cost and profit of the businesses. For example, if the oil price increases, the stock price will decrease due to the business's higher cost and lower profit ([15, 26]). Second is the inflation from the stock market driven by changes in oil prices. The higher inflation will decrease money's value, thereby leading to the stock market downturn [18]. Thirdly, the higher inflation due to the oil price in sometimes can be solved by using interest rate policy. However, this policy reduces a firm predicted present cash flow's value, which lowers the value of their stock [9]. Many empirical studies have also confirmed the negative impact of oil prices on stock markets ([5, 10, 11]). In a study by [10], it was discovered that changes in energy costs have a long-term association with the stock market. This relationship can be seen when comparing oil price data with the MSCI World Index (an index that measures overall global stock performance).

This study analyzes the impact of both financial and health crisis events and oil price on the Japanese market and compares the exposure extent. The unusual situation of the Japanese stock market is the primary emphasis of this study. In comparison to other countries, the impact of the COVID-19 spread in Japan has been moderate [25]. However, the Japanese stock market dropped by around 30% decline during February to March of 2020, which was almost the same as that of the US. Likewise, the Japanese market also experienced an approximately 60% drop during 2007–2009, which was also similar to the US (origin of the crisis). Although Japan is not the origin of both crises, it is quite surprising that Japanese stock has a large drop after the crisis occurred. This phenomenon is not well explained by the health and financial crisis impacts on the Japanese stock market.

Our paper has some novelties first; although the origin and dynamics of both crisis events and oil price movement are different, the effects of these factors on the financial market are comparable because of the similarities in financial transmission and spillovers [13]. The study offers an extensive examination of

both health and financial crisis events and oil prices regarding their impacts on the Japanese stock market. Second, by employing the nonlinear autoregressive distributed lag model (ARDL), the current study answers an important research question: what are the key risk factors that can abate the Japanese market return? Third, the majority of earlier studies focused at how oil price shocks symmetrically affected stock market returns [6, 7]; this study explores the asymmetric impact of oil shocks (positive and negative shocks) on stock return.

The rest of the paper is organized as follows. Data and methodology are provided in Sect. 2. Empirical results are described and analyzed in Sect. 3, and Sect. 4 presents conclusions.

## 2 Methodology

This study investigates the impacts of major epidemics, financial crises, and oil price shocks in isolation on the Japanese stock market in terms of daily market return proxied by the NIKKEI 225 Index (JAP). As we are dealing with the asymmetric oil price, the Brent oil price (OP) is also collected and decomposed into two new time series variables using partial sum concepts [24], named positive oil shock ( $OP^+$ ) and negative oil price shock ( $OP^-$ ) as follows

$$OP_t^+ = \sum_{i=1}^t \max(\Delta OP_i, 0), \tag{1}$$

$$OP_t^- = \sum_{i=1}^t \min(\Delta OP_i, 0) \tag{2}$$

where  $\Delta OP$  is the differentiation between oil prices at the time  $t$  and  $t-1$

Regarding the methodology, there are several advantages to using a nonlinear or asymmetric ARDL model. First, if the essential variables have a mixed order of integration or are not entirely non-stationary, it can be implemented directly [8]. Second, it enables us to examine oil prices' dynamic and asymmetric effects on Japanese stock returns. Third, the model considers the short and long-run effects of independent variables on Japanese stock return [24].

From formula 3, JAP is the NIKKEI 225 Index, MSCI is the MSCI World index return which is used as the control variable (an index that measures overall global stock performance), and the two dummy variables used are financial (FC) and health (HC) crises. We can write our empirical nonlinear ARDL model as the following

$$\begin{aligned} \Delta JAP_t = & \alpha + \sum_{i=1}^{p1} \beta_{1i} \Delta JAP_{t-i} + \sum_{i=1}^{p2} \beta_{2i} \Delta MSCI_{t-i} + \sum_{i=1}^{p3} \beta_{3i} \Delta OP_{t-i}^+ + \sum_{i=1}^{p4} \beta_{4i} \Delta \\ & OP_{t-i}^- + \sum_{i=1}^{p5} \beta_{5i} \Delta FC_{t-i} + \sum_{i=1}^{p6} \beta_{6i} \Delta HC_{t-i} + \rho_1 JAP_{t-1} + \rho_2 MSCI_{t-1} \\ & + \rho_3 OP_{t-1}^+ + \rho_4 OP_{t-1}^- + u_t \end{aligned} \tag{3}$$

where  $\alpha$ ,  $\beta$  and  $\rho$  represent intercept term, short-run coefficient and long-run coefficients, respectively. is the error term. Note that  $\beta_{3i}$  and  $\rho_3$  can be viewed as the positive shock of oil price in the short-run and long-run, respectively. While  $\beta_{4i}$  and  $\rho_4$  indicate the positive oil price shock impact. [24] refer to it as a nonlinear ARDL model since the construction of the  $OP^+$  and  $OP^-$  variables introduces nonlinearity into the adjustment process in (3).

In addition, to check the integrated order of the variables before applying this model, it is necessary to check the cointegration of the variables using the bound test proposed by [23]. They provide F-test statistics' lower and upper bounds. The acceptance of the null hypothesis and conclusion that there is no cointegration has occurred if the F-test statistics fall below the lower bound critical value. However, we reject the null hypothesis if the F-test statistics are higher than the upper bound value and conclusion that cointegration has occurred. When the F-test statistic occurs between the lower and upper bound critical values, the test is inconclusive.  $\rho_1, \rho_2, \rho_3, \rho_4, \rho_5$  are long-run coefficients of Equation (3), the null hypothesis of the bound test set all of these coefficients are equal to zero ( $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = 0$ ). Moreover, determining an appropriate lag for the model is another important issue, this study considered the Akaike Information Criterion (AIC), and the lowest AIC is preferred.

### 3 Data Description

This daily data is collected covering January 2005 to March 2022. All data is obtained from the investing.com database. This study considers two dummy variables that represent financial (FC) and health (HC) crises regarding the financial and health crises. The first dummy variable is for health crises, taking a value of one during the event periods (Bird flu (2005), Swine flu (2009), Ebola (2014) and the COVID-19 pandemics (2020–2022) and zero otherwise and the second dummy is for financial crisis covering the subprime mortgage crisis (2008–2009) and the European sovereign debt (2012). In addition, to avoid the omitted variable, we add the MSCI World index return (MSCI) as a control variable. The descriptive statistics of each variable is presented as follows:

According to Table 1, the daily means of JAP and MSCI are 0.02 %. However, the standard deviation of the JAP is larger than MSCI, indicating the higher fluctuation of Japanese stock compared to the global stock. Considering the core independent variables, the average positive shock and negative oil price are  $-0.79\%$  and  $0.81\%$ , respectively, implying that the positive shock is larger than the negative shock.

Finally, the mean of FC is higher than HC, indicating that the financial crisis occurs more frequently than the health crisis during 2005–2022.



**Table 1.** Descriptive Statistics

	JAP	MSCI	$OP^-$	$OP^+$	FC	HC
Mean	0.0002	0.0002	-0.0079	0.0081	0.5366	0.4211
Median	0.0005	0.0006	0.0000	0.0005	1.0000	0.0000
Maximum	0.1323	0.0909	0.0000	0.4120	1.0000	1.0000
Minimum	-0.1211	-0.1044	-0.6436	0.0000	0.0000	0.0000
Std. Dev	0.0146	0.0106	0.0184	0.0161	0.4987	0.4938
Skewness	-0.4447	-0.7647	-12.980	7.8198	-0.1468	0.3194
Kurtosis	10.702	15.882	364.17	137.67	1.0215	1.1020
Jarque-Bera	10497	29390.22	22897	32098	698.58	700.31
Probability	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

## 4 Results

### 4.1 Unit Root Test

The Augmented Dickey-Fuller (ADF) test is used to check the integrated order of each variable. the variable is integrated at zero  $I(0)$  when the variable is stationary at level. If the variable is stationary at the first difference, the variable is integrated at one  $I(1)$ . The result of the ADF test shown in Table 2. We can observe that JAP,  $OP^-$ , and  $OP^+$  are integrated at zero, while MSCI is integrated at one. For HC and FC, we are not testing their integrated order as both of them are dummy variables

**Table 2.** Unit root test Result

	JAP	MSCI	$OP^-$	$OP^+$
ADF test (level)	-3.097	-0.402	-17.695	-28.235
(MBF)	0.008	0.922	0.000	0.000
ADF test (1st diff)		-61.015		
(MBF)		0.000		
Integrated order	I(0)	I(1)	I(0)	I(0)

Note: The ADF-test and Jarque-Bera test are inferred using the Maximum Bayes factor (MBF). (see, [20])

### 4.2 The Cointegration Result

Table 3 reports the cointegration result from the bound test. In this section, four MBF criteria are considered to examine the cointegration between the variables.

**Table 3.** Bound Test

Test Statistic	Value	MBF evidence	Lower	Upper
F-statistic for Japan	2.9413	moderate	1.92	2.89
		strong	2.17	3.21
		Very strong	2.43	3.51
		decisive	2.73	3.90

**Table 4.** Short run asymmetry

Variable	Coefficient	Std. Error	t-statistic	MBF
$\Delta JAP_{t-1}$	-0.0336	0.0154	-2.1849	0.0910*
$\Delta JAP_{t-2}$	0.0143	0.0154	0.9299	0.6489
$\Delta JAP_{t-3}$	-0.0285	0.0144	-1.8538	0.1793
$\Delta MSCI_{t-1}$	0.8392	0.0283	29.6537	0.0000*
$\Delta FC_{t-1}$	-0.0015	0.0003	5.2274	0.0000*
$\Delta HC_{t-1}$	-0.0024	0.0009	-2.6666	0.0023*
$\Delta OP^+_{t-1}$	-0.0008	0.0002	-2.9345	0.0134*
$\Delta OP^-_{t-1}$	-0.0006	0.0002	-2.8239	0.0185*

Note: (\*) indicates strong evidence supporting the effect of this variable

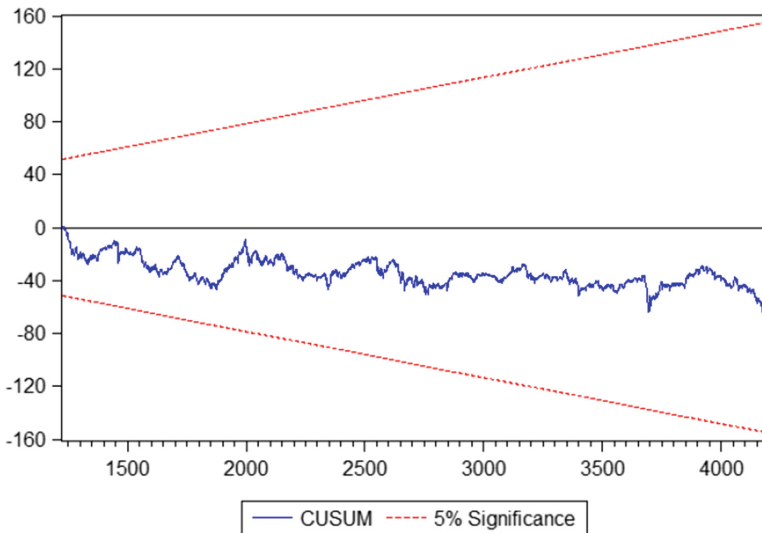
**Table 5.** Long run asymmetry

Variable	Coefficient	Std. Error	t-Statistic	MBF
$MSCI_t$	1.1574	0.1602	7.2244	0.0000*
$FC$	-0.1039	0.1195	-0.8694	0.6852
$HC$	-0.1154	0.1125	1.0261	0.5906
$OP^-$	-0.0765	0.0925	-0.8272	0.7102
$OP^+$	0.0681	0.0930	0.7325	0.7646
$\alpha$	1.1301	0.0534	21.162	0.0000*

Note: (\*) indicates strong evidence supporting the effect of this variable

**Table 6.** Results of diagnostic tests

	$\chi^2$ statistic	MBF	t-Statistic	MBF
Breusch-Godfrey Serial Correlation test	0.5136	0.7735	7.2244	0.0000*
White Heteroskedasticity test	7.5753	0.8728	-0.8694	0.6852
Jarque-Bera test	0.8925	0.5128	1.0261	0.5906



**Fig. 1.** The plot of the CUSUM test

The result shows that the null hypothesis of no cointegration is rejected as the F-statistic value falls between the lower bound and upper bound of the bound test with decisive evidence. Our evidence confirms the long-run relationship between Japanese stock returns and oil, health crisis, and financial crisis.

After determining the cointegration relationships, we then report the short and long-run coefficients in Tables 4 and 5, respectively.

Considering the short-run effects of a health crisis, financial crisis, and oil price shocks in Table 4, our results show a negative impact of financial and health crises with strong evidence. Comparing these two crises, we can observe that the health crisis is likely to contribute a larger impact than the financial crisis, which is in line with the literature [13]. The possible reason is that the impact of many health crisis events is spreading worldwide, but the financial crisis is sometimes limited in the region. In terms of oil price shocks, the literature indicates that oil prices have asymmetric and considerable effects on the stock market ([9];[19]). Different positive and negative volatility in oil prices will confirm the asymmetry. Although both positive and negative oil price shocks negatively affect the Japanese stock return, the magnitude of these two crises is different, in which the impact of the positive shock is larger than the negative. [9] explained that a higher oil price could lead to a substantial increase in inflation, thereby lowering the expected cash flow of a firm. Moreover, we gather that the Japanese stock return carries at least one lagged significant coefficient as the MBF value is less than 0.1. This implies that stock return on the previous day has a short-run effect on the current stock return.

We then move to the result of long-run effects. The result is presented in Table 5. The findings suggest that there are weak evidence effects of all variables

on the stock return, except for MSCI. This indicates that the short-run effects of health crises, financial crises, and oil price shocks do not last into long-run effects.

After the nonlinear ARDL test results, the diagnostic tests are adopted to examine the goodness of fit of our model. These include serial correlation, heteroscedasticity, and normality of errors. The model also passes all the diagnostic tests, and all the results are shown in Table 6. Moreover, The CUSUM test also confirms that our model satisfies the stability conditions. Because there is no root outside the significance level (See in Fig. 1), thus, it can be concluded that the parameters of the nonlinear ARDL on Japanese stock return are stable.

## 5 Conclusions

This study aims at finding the longrun and short-run impacts of financial and health crises and oil price shocks on the Japanese stock market. For empirical analysis, we employ the nonlinear ARDL. This study's main finding shows that the asymmetric of Japanese stock market has occurred in the short run, as the inverse impact of a positive oil shock is larger than the negative price shock. As Japan is a large oil-importing nation, oil becomes the main cost of production. Therefore, changes in oil prices are the main factor affecting the stock market in the short run [19]. We also find negative influences of health and financial crisis on the return of the stock, in which the impact of a health crisis is larger than the financial crisis. We expect that critical factors such as the number of cases or deaths may be more contributing factors, causing the stock market to decline [2]. However, the link between the financial crisis and health crisis, oil, and stock return of Japan in long term does not exist.

For further study, the dynamic causality between oil price shock and stock is suggested in order to obtain new insight [21]. Moreover, the scope of this study can be expanded to other countries in order to check the robust impacts of financial and health crises and oil prices.

**Acknowledgments.** The authors are grateful to the Centre of Excellence in Econometrics, Chiang Mai University, for financial support. The authors are grateful to Dr. Laxmi Worachai for her helpful comments and suggestions.

## References

1. Al Refai, H., Zeitun, R., Eissa, M.A.A.: Impact of global health crisis and oil price shocks on stock markets in the GCC. *Financ. Res. Lett.* **45**, 102130 (2022)
2. Alamgir, F., Amin, S.B.: The nexus between oil price and stock market: evidence from South Asia. *Energy Rep.* **7**, 693–703 (2021)
3. Ali, B., Khan, D., Shafiq, M., Magda, R., Oláh, J.: The asymmetric impact of oil price shocks on sectoral returns in Pakistan: evidence from the nonlinear ARDL approach. *Economies* **10**(2), 46 (2022)
4. Ali, R., Afzal, M.: Impact of global financial crisis on stock markets: evidence from Pakistan and India. *J. Bus. Manage. Econ.* **3**(7), 275–282 (2012)

5. Antonakakis, N., Chatziantoniou, I., Filis, G.: Oil shocks and stock markets: dynamic connectedness under the prism of recent geopolitical and economic unrest. *Int. Rev. Financ. Anal.* **50**, 1–26 (2017)
6. Asaad, Z.: Oil price, gold price, exchange rate and stock market in Iraq pre-during COVID-19 outbreak: an ARDL approach. *Int. J. Energy Econ. Policy* **11**(5), 562–671 (2021)
7. Atri, H., Kouki, S., Imen Gallali, M.: The impact of COVID-19 news, panic and media coverage on the oil and gold prices: an ARDL approach. *Resour. Policy* **72**, 102061 (2021)
8. Banerjee, A., Dolado, J.J., Galbraith, J.W., Hendry, D.: *Co-integration, Error Correction, and the Econometric Analysis of Non-stationary Data*. Oxford University Press, Oxford (1993)
9. Civcir, I., Akkoc, U.: Nonlinear ARDL approach to the oil-stock nexus: detailed sectoral analysis of the Turkish stock market. *Resour. Policy* **74**, 102424 (2021)
10. Cong, R. G., Shen, S.: Relationships among energy price shocks, stock market, and the macroeconomy: evidence from China. *The Scientific World Journal* (2013)
11. Degiannakis, S., Filis, G., Kizys, R.: The effects of oil price shocks on stock market volatility: Evidence from European data. *Energy J.* **35**(1) (2014)
12. Ganie, I.R., Wani, T.A., Yadav, M.P.: Impact of COVID-19 outbreak on the stock market: an evidence from select economies. *Bus. Perspect. Res.* 22785337211073635 (2022)
13. Gunay, S., Can, G.: The source of financial contagion and spillovers: an evaluation of the COVID-19 pandemic and the global financial crisis. *PLoS ONE* **17**(1), e0261835 (2022)
14. Jebran, K., Chen, S., Ullah, I., Mirza, S.S.: Does volatility spillover among stock markets varies from normal to turbulent periods? Evidence from emerging markets of Asia. *J. Finance Data Sci.* **3**(1–4), 20–30 (2017)
15. Jones, C.M., Kaul, G.: Oil and the stock markets. *J. Financ.* **51**(2), 463–491 (1996)
16. Karpunina, E.K., Moskovtceva, L.V., Zabelina, O.V., Zubareva, N.N., Tsykora, A.V.: Socio-economic impact of the COVID-19 pandemic on OECD countries. In: *Current Problems of the World Economy and International Trade*, Emerald Publishing Limited (2022)
17. Kenton, W.: Financial sector (2021). [https://www.investopedia.com/terms/f/financial\\_sector.asp](https://www.investopedia.com/terms/f/financial_sector.asp). Accessed 24 Aug 2022
18. Kilian, L., Park, C.: The impact of oil price shocks on the US stock market. *Int. Econ. Rev.* **50**(4), 1267–1287 (2009)
19. Le, T.H., Chang, Y.: Effects of oil price shocks on the stock market performance: do nature of shocks and economies matter? *Energy Econ.* **51**, 261–274 (2015)
20. Maneejuk, P., Yamaka, W.: Significance test for linear regression: how to test without P-values? *J. Appl. Stat.* **48**(5), 827–845 (2021)
21. Maneejuk, P., Yamaka, W., Sriboonchitta, S.: Entropy inference in smooth transition kink regression. *Commun. Stat. Simul. Comput.* **51**, 7366–7389 (2020)
22. Pastpipatkul, P., Yamaka, W., Wiboonpongse, A., Sriboonchitta, S.: Spillovers of quantitative easing on financial markets of Thailand, Indonesia, and the Philippines. In: Huynh, V.-N., Inuiguchi, M., Denoeux, T. (eds.) *IUKM 2015. LNCS (LNAI)*, vol. 9376, pp. 374–388. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-25135-6\\_35](https://doi.org/10.1007/978-3-319-25135-6_35)
23. Pesaran, M.H., Shin, Y., Smith, R.J.: Bounds testing approaches to the analysis of level relationships. *J. Appl. Economet.* **16**(3), 289–326 (2001)

24. Shin, Y., Yu, B., Greenwood-Nimmo, M.: Modelling asymmetric cointegration and dynamic multipliers in a nonlinear ARDL framework. In: Sickles, R.C., Horrace, W.C. (eds.) *Festschrift in Honor of Peter Schmidt*, pp. 281–314. Springer, New York (2014). [https://doi.org/10.1007/978-1-4899-8008-3\\_9](https://doi.org/10.1007/978-1-4899-8008-3_9)
25. Takahashi, H., Yamada, K.: When the Japanese stock market meets COVID-19: impact of ownership, China and US exposure, and ESG channels. *Int. Rev. Financ. Anal.* **74**, 101670 (2021)
26. Vätavu, S., Lobonț, O.R., Para, I., Pelin, A.: Addressing oil price changes through business profitability in oil and gas industry in the United Kingdom. *PLoS ONE* **13**(6), e0199100 (2018)



# Tourism Business Adaption to Survive the Coronavirus Disease-2019 Pandemic in Thailand

Supareuk Tarapituxwong<sup>1</sup>, Piangtawan Polard<sup>1</sup>, and Namchok Chimprang<sup>2</sup>(✉)

<sup>1</sup> Faculty of Management Sciences, Chiang Mai Rajabhat University,  
Chiang Mai 50300, Thailand

{supareuk\_tar,piangtawan\_pol}@cmru.ac.th

<sup>2</sup> Center of Excellence in Econometrics, Faculty of Economics,  
Chiang Mai University, Chiang Mai, Thailand  
namchok.c@outlook.co.th

**Abstract.** The tourism industry is a service sector playing a significant role in the Thai economy, accounting for 11 percent of the total revenue generated domestically before the coronavirus 2019 pandemic (COVID-19). Unfortunately, since some COVID-19 countermeasures like lockdowns and travel restrictions have drastically reduced the number of tourists, this revenue continuously diminishes. Consequently, many tourism-related businesses and activities could not survive through the circumstances. The present research aims to investigate the supporting and inhibiting factors of tourism business survival during the coronavirus pandemic, firstly, using the Cox proportional hazards model. Then, the Kaplan-Meier estimator is employed to estimate the businesses' survival probability of individual significant parameter across survival periods. The results reveal that the survivability of a tourism-related establishment mainly depends on firm's characteristics such as business type, location, fixed assets, and net income. Meanwhile, such business and employment strategies as social distancing practice, laying off part-time employees, and operating without working-hour reduction can ameliorate the business's survivability. Furthermore, domestic-tourist targeting enterprises appear to contribute to high survivability similar to the exporting and importing components in their supply chain. Among various government financial relief measures, the reduction of contribution of employers to the Social Security Fund is found to be a supporting factor for business survivability. However, some debt-relieving measures worsen the probability of business survival. Lastly, the survival probability paths over the studied period from the Kaplan-Meier estimation show that any business adaptation including government support should be promptly employed within the first six months or at the latest within a year of this epidemic event to increase the survivability.

**Keywords:** Business survival · Tourism · Coronavirus pandemic · Cox proportional hazards model, Thailand

## 1 Introduction

The tourism sector in Thailand possesses a considerable potential for development into a global tourism hub. The top 10 visiting and earnings from tourism countries report by the United Nations World Tourism Organization (UNWTO) demonstrates the potential of Thailand's tourism industry to be more outstanding than other countries in the ASEAN region [1]. The tourism industry as a service sector had played a significant role in the Thai economy until the arrival of COVID-19 because its revenue in 2019 (2 trillion baht) accounts for 11 percent of the Thai GDP. Unfortunately, this revenue is likely to diminish (6.1 hundred billion baht) with the estimated foreign tourist numbers below five hundred thousand in 2020–2021 [2]. In Thailand, several enterprises are involved in the tourism industry, including hotel business, travel agencies, tour guides, transportation, food retail, beverage, souvenirs, entertainment, nature and rural tourism, etc. Therefore, the sustained expansion of the tourism industry will essentially contribute to the continuous growth of other business sectors. However, Thailand's tourism still faces a bunch of issues, most notably a decline in tourism revenue due to the coronavirus pandemic.

The announcement of the presence of the coronavirus pandemic in January 2020 triggered tremendous economic turmoil, particularly in developing countries, as well as the failure of numerous firms around the world [3]. Thailand's economy also has received a severe impact like other nations. The tourism sector especially has witnessed an unprecedented contraction since the implementation of complete country lockdown and travel restriction measures around the middle of 2020. Many business establishments have permanently closed since they could not overcome the economic contractions brought about by these radical measures [4]. However, some tourism-related enterprises can survive through the COVID crisis with their adaptation strategies and operational characteristics. In Thailand, the government has formulated several economic stimulus policies and countermeasures against the severe spread of the coronavirus to restore the economic health that has been devastated by the pandemic. The easing of the country's lockdown, the We Travel Together scheme, the Half-Half project, and the provision of extended public holidays have all been developed as part of its economic stimulus initiatives to encourage spending and domestic tourism. Besides, if the coronavirus pandemic scenario has subsided along with the recovery of the tourism sector, the government must promptly restructure the tourism sector in order to accommodate the new normal behavior of tourists. Especially, tourism entrepreneurs will have to adjust their business strategies to reflect the shifting traveler preferences around the globe, emphasizing health and hygiene, avoiding crowds, and concentrating more on niche travel than mass travel.

Therefore, a study of the surviving tourism entrepreneurs along with the failed ones during the coronavirus pandemic will serve as a guide for improving business competitiveness. Previous studies have demonstrated the significant effects of the coronavirus pandemic in several aspects, such as on society, the economy, and the environment. Furthermore, many studies suggest that financial performance [5], resilience strategies and innovation [6], and geographic location



[7] should be factors in assessing the coronavirus pandemic impact on the social enterprise. According to Shafi et al. [8], most companies were impacted by the crisis severely and were struggling with many challenges, including mounting debt, a tightened supply chain, declining demand, and profit loss. Likewise, Huynh et al. [9] found that the degree of the business downturn from the coronavirus impact differed among the different tourism enterprise types. Ma and Gao [10] explain that government policies can indirectly lessen the pandemic impact on SMEs. However, the direct influence of the pandemic on the market is the most noteworthy. Furthermore, Gregurec et al. [11] reveal that several countries have implemented lockdown measures and halted corporate operations to prevent the spread of coronavirus. Numerous enterprises are thus temporarily suspended or permanently shut down.

In this study, we focus on studying the probability of business failure given significant risk factors such as firm's characteristics, ongoing crisis-related business adaption, target customer, supply chain, employment, and government support policies. These factors are essential for developing and modifying appropriate policies to support the businesses during this crisis. To accomplish the goal, the Cox proportional hazards model [12] and the nonparametric Kaplan-Meier estimator [13] are used in the survival study of the Thai tourism-related enterprises. Overall, our results can provide specific guidance for tourism businesses and the government on methods for strengthening the survival probability of the businesses of interest and how to deal with the negative impact of the pandemic during the COVID-19 situation.

The remainder of this study is structured as follows: Sect. 2 gives details on the methodology, Sect. 3 describes the data used, Sect. 4 presents the empirical findings, and Sect. 5 provides the conclusion and policy implications.

## 2 Methodology

This section provides a background on the two-step approach used in the study. First, we discuss the theory underlying Cox proportional hazards modeling and survival analysis in general. The Cox proportional hazards model is considered to be particularly effective at identifying the risk factors affecting the probability of a Thai firm's survival. Since we cannot directly interpret the coefficient estimates from this Cox model, hazard ratios (HR) of each coefficient are computed. Second, the firm's probability of surviving for the coming 12 months is then investigated, considering the influence of individual variables. The firm's survival path can be obtained and illustrated by using the nonparametric Kaplan-Meier estimator [12].

### 2.1 Survival and Hazard Functions

Let the survival time  $T$  be a random variable with cumulative distribution function  $P(t) = \Pr(T \leq t)$ , and probability density function  $p(t) = dP(t)/dt$ . The survival function is the complement of the distribution function,  $S(t) =$

$\Pr(T > t) = 1 - P(t)$ . And the relationship between  $S(t)$  and  $h(t)$  the hazard function, which estimates the sudden risk of failure at time  $t$ , is expressed by:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)} \tag{1}$$

where  $f(t) = d(S(t))/d(t)$  is the probability density of failure at time  $t$ . This implies that

$$h(t) = \frac{dS(t)/d(t)}{S(t)} = \frac{d}{dt}(-\log S(t)), \tag{2}$$

or equivalently,

$$S(t) = e^{-H(t)}, \tag{3}$$

where  $H(t)$  is the cumulative hazard function. This function indicates the probability that an individual will experience an event (for example, death or bankruptcy) within a considered time period [14].

**2.2 The Cox Proportional Hazards Model**

Survival analysis typically addresses the relationship of the survival distribution with covariates. The simplest way to examine this relationship is the linear regression specification.

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}, \tag{4}$$

or, again equivalently,

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}), \tag{5}$$

where  $h_i(t)$  is the conditional hazard function of continuous random variable. Briefly, the hazard function can be interpreted as the failure of business  $i$  and the survival time  $t$ . If business  $i$  could not survive during the crisis,  $h_i$  is given as 1, otherwise  $h_i = 0$ .  $k$  is the number of independent variables,  $\beta_i$  are the partial regression coefficients,  $x_i$  are the independent variable of individual observation  $i$ ,  $h_0(t)$  is the hazard baseline. If the value of  $\beta_i$  is greater than zero, or equivalently a hazard ratio greater than one, it indicates that the event's hazard increases, and the length of survival will decrease. we illustrate the concept in Fig. 1

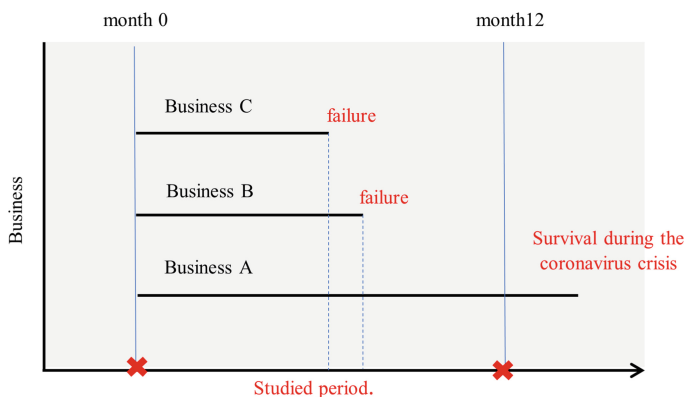


Fig. 1. The illustration of business survivability

### 2.3 Kaplan-Meier Estimator

The Kaplan-Meier estimation is one of the most popular methods of computing survival over time conditional on the predictor. Specifically, only one predictor is considered by the Kaplan-Meier estimate to construct the survival curve, which is defined as the probability of surviving over a specified period. This estimator is a nonparametric one with a few restrictions [15] and the properties that both the event of interest and the study period are clearly defined, and the survival probability of all businesses and the censored observations is the same. In our analysis, only significant factors obtained from the Cox regression are used to illustrate the survival probability predicting curve during this coronavirus crisis by the Kaplan-Meier estimator.

## 3 Data Specification

This study examines the survivability factors of Thai tourism enterprises during the coronavirus pandemic using a survey derived from part of the Asia Foundation’s revisiting the pandemic project. The project gathered new information on the state of the Thai tourism enterprises during the coronavirus pandemic in December 2021. The survey also collects information on six different aspects of the Thai tourism business, including firm’s characteristics, ongoing crisis-related business adaption, target customer, supply chain, employment, and government support policies. These aspects are considered independent variables in the study. Information concerning each part is displayed in Table 1. The survey samples include 400 Thai tourism-serving firms collected equally from Thailand’s four major geographic regions: the North, Northeast, Central, and South. However, there are differences in the sample distribution per province within each area. To determine the economic survival probability of Thai tourism businesses, information related to the survival period is required to assess the possibility that the establishments will remain in operation economically if Thailand has another

year of the coronavirus crisis. The definitions of the variables used in this study are presented in Table 1. Note that the average is presented for continuous variables while corresponding percentages (%) are used for others. According to Table 1, we observe that percentage of non-survival businesses from this crisis is 33.5%, while the survival period average is 9.2 months.

Considering the independent variables, the majority 24.00% of the tourism-serving establishments are of the food and beverage type, and the least in number is the transportation type with 8.25%. Most of the firms surveyed, 42.25%, indicated earning net income between THB 1-5 million, while 36.75% had total fixed assets of THB 3-10 million, and 62.75% operated on Rent/Lease business premises. Furthermore, 93.50% of the businesses under study reported facing a business revenue loss by 54.53% as the result of the coronavirus pandemic.

In terms of business adaptation, more than 60% of the tourism-related firms currently operate as usual with regular working hours, while 41.00% operate with shortened working hours during the first wave of coronavirus in March 2020. For the business employment factor, we find that tourism businesses have an average of 44.97 persons for full-time employees and 11.75 persons for part-time employees. To survive this wave of crisis, 46.5% of the businesses laid off their employees. 23.25% reduced working hours to minimize the layoff of employees, and 24.75% continued doing everything as usual. If we look at the number of layoffs due to the coronavirus pandemic, the average is 10.26, 8.06, and 1.27 persons in the category of permanent employee, part-time employee, and expected laid-off employee in the next two months, respectively. Interestingly, less than half of the tourism businesses had logistics and supply chains components involving importing or exporting goods or services. In other words, most tourism-related establishments play a significant role in employment and the local economy of Thailand.

In the case of government support policies, the measure to reduce contributions to the Social Security Fund for employees and employers was most popularly participated by the tourism related businesses followed by the measures to stimulate domestic tourism by adding benefits for registrants of the “We travel together” scheme.

## 4 Empirical Results

### 4.1 Estimation Results of the Cox Proportional Hazards Model

Table 2 presents the obtained coefficients from the Cox regression model using the non-penalized method. Unfortunately, we cannot directly interpret the coefficients because each independent variable’s effect on the business failure varies. Therefore, the exponential is used for converting these coefficients into hazard ratios. If the hazard ratio value is greater than 1, the factor will explicitly result in a lower survival probability for the business than the reference group. Contrarily, the hazard ratio value below 1 indicates that the factor increases the survival probability of the business compared with the reference group [16].

**Table 1.** Descriptive statistics

Variable	Description	Percentage
<b>Dependent variable: Survival data</b>		
Business failure	Business failure within 12 months	33.50
	Business survives more than 12 months	66.50
Economic survival duration	Number of month that business can operate	9.20
<b>Independent variable: Fundamental business factors</b>		
Types of business related to tourism sector	Hotel / accommodations	21.25
	Transportation	8.25
	Souvenir	20.50
	Food and beverage	24.00
	Travel agency / tour guide	17.50
	Other business in tourism sector	8.50
Total fixed assets	Total fixed assets less than THB 3 million	13.00
	Total fixed assets THB 3–10 million	36.75
	Total fixed assets THB 11–20 million	26.25
	Total fixed assets more than THB 20 million	24.00
Net income (Annual)	Net income less than THB 1 million	7.75
	Net income THB 1–5 million	42.25
	Net income THB 6–10 million	30.50
	Net income more than THB 10 million	19.50
Location of business	Central region	25.00
	Eastern region	25.00
	Northern region	25.00
	Southern region	25.00
Business premises	Does not have/ use a business premises	11.00
	Rent/Lease a business premises	62.75
	Own a business premises	26.25
Decreased revenue/income change	The percentage change in the business's revenue during the coronavirus pandemic. Decreased revenue change is presented by a positive value.	54.53*
<b>Independent variable: Epidemic and business adaptation factors</b>		
Perceiving the risks of the coronavirus impact	Perceiving the risks of the coronavirus impact of entrepreneur is divided into 5 levels of awareness.	3.31*
The current operation of a business	Open as usual and regular working hours	68.50
	Open with shortened working hours and days	25.00
	Not accepting customers but the employees are still working such as work from home and delivery.	4.00
	The firm is closed and has not yet reopened	2.50
The operation of a business during the first wave of the coronavirus pandemic in March 2020	Open as usual and regular working hours	26.25
	Open with shortened working hours and days	41.00
	Not accepting customers but the employees are still working such as work from home and delivery	32.75
Business model change from the coronavirus pandemic.	Operate while adapting to social distancing	74.00
	Move into new products and services that have high demand during the coronavirus pandemic	43.25
	Operate through online markets or social media	64.50
	Discussed with employees to reduce their salary to keep all employees	23.50
	No adaption	13.00

(continued)

**Table 1.** (continued)

Variable	Description	Percentage
<b>Independent variable: Employment factors</b>		
Number of employees	Full-time employees	44.97*
	Part-time employees	11.75*
Dismissal of employees	Number of laid-off permanent employees	10.26*
	Number of laid-off part time employees	8.06*
	Number of employees expected to be laid off in the next 2 months	1.27*
Labor Management	Reduced working hours to lower the unemployment rate	23.25
	Doing everything as usual	24.75
	Lay off employees (Part/All)	46.50
	Temporarily closed	5.50
<b>Independent variable: Target market factor</b>		
Target customer	Domestic tourists' ratio	62.31*
Change in the number of tourists	At the same level as before the coronavirus pandemic	0.50
	Less than 25% reduction	7.50
	Reduced by about 25–50%	46.50
	Decreased by more than 50%	43.00
	No domestic tourists	1.00
	Increased more than usual	1.50
<b>Independent variable: Logistics and Supply Chain Factors</b>		
Import/Export of goods and services	Import	15.25
	Export	7.75
	Import and export	12.00
	No import and export	65.00
<b>Independent variable: government support policies</b>		
Government support policies during the pandemic	Low-interest loan measures, Soft Loan, 2% interest, and a 6-month moratorium on debt.	11.75
	Tax deductible measure for businesses that still employ all employees.	11.75
	Measures to reduce contributions to the Social Security Fund for employees and employers	64.50
	Relaxation measure for personal and business tax filing	26.25
	Measures to reduce withholding tax rates from services	23.50
	Measures to extend the period of payment of income taxes (such as VAT) and excise taxes	19.00
	Measures to stimulate domestic tourism by adding benefits for registrants “We travel together”	57.25
	Not participating in any project	11.50

**Note:** The star (\*) indicates the average value of continuous variable.

We present the estimated coefficients with their hazard ratios in Table 2. Our results reveal that the fundamental characteristics factors of a tourism business consisted of business type, total fixed assets, net profit (per year), business location, business premises (region), and percent change in revenue statistically significantly impacted the probability of business survival. The food and beverage

business type gets the most negative impact on business failure corresponding to the hazard ratio of 0.0944, indicating that the food and beverage business has the lowest risk of non-survival than other business types. Less total fixed assets businesses are less prone to failure than those with more. According to Table 2, findings show that total fixed assets of less than 3 million baht and between 3 to 10 million baht have a negative impact on the failure of a business, with hazard ratios of 0.2479 and 0.5862, respectively. On the other hand, lower annual net income businesses appear to have a higher possibility of failure. As expressed in Table 2, businesses with net incomes of less than 1 million baht and between 1 and 5 million baht have a failure possibility of 8.1390 and 3.1040 times, respectively, compared to more than those 10-million-baht net income businesses. Furthermore, businesses in the South have a negative impact on business failure corresponding to the hazard ratio of 0.4388, indicating a lower risk of non-survival than those in the Central region. Furthermore, a decrease in revenue of a business by 1% worsens the business's survival probability by 3.70%.

Concerning business strategies factors for surviving the pandemic, we find that if the perceived risks of the coronavirus impact increase 1 level, its non-survival probability increases by 2.0803 times, indicating accurate risk awareness of the business. Therefore, the government can employ this factor as an indicator to monitor high-risk businesses and then effectively provide them support or help. Moreover, adapting the business model to accommodate social distancing leads to an increase in the possibility of survivability by 59.62%.

We then discover that increasing the number of laid-off part-time employees ameliorates the business's survival probability. The higher number of laid-off results in survival probability increase by 5.55%, corresponding to the hazard ratio of 0.9450. Moreover, the regular operation using employees without working hours reduction also increases the business survival probability by 74.84%.

In terms of target market factors, we find a negative effect of the domestic visitors' proportion on business survivability as 1 level of domestic tourist ratio growth can increase a firm's possibility of survival by 30.87%. On the other hand, a decrease in the number of tourists by more than a quarter can severely worsen the business's survival probability with a hazard ratio of more than 9.8340 compared to a normal situation. Furthermore, the result shows that exporting and importing in business supply chain and logistics have a greater probability of surviving 92.33% (0.0767 hazard ratio) than those without these actions.

Lastly, we observe that two government-supporting policies are the key risk factors affecting the survival of Thai tourism-related enterprises. First, the debt-involving measures, comprised of Low-interest loans, Soft Loan, 2% interest, and a 6-month moratorium on debt measures, decisively worsen the probability of business survival, with a hazard ratio of 6.8636 compared to not participating in the scheme. As a result, these measures can slow down the business bankruptcy but might not help business to survive over the long term. On the contrary,

reduction of contributions to the Social Security Fund is a supporting factor of a business’s survival by decisively increasing the probability of survivability by about 74.37% compared to not participating in this scheme.

**Table 2.** Results of the Cox regression model

Variable	Coefficient	Hazard ratio	SE	Z	MBF
<i>Fundamental business factors</i>					
<b>Business type</b>					
Hotel	-1.6857 **	0.1853	0.6793	-2.482	0.0460
Transportation	-0.6822	0.5055	0.7207	-0.947	0.6389
Souvenir	-1.9746 **	0.1388	0.8467	-2.332	0.0659
Food and beverage	-2.3597 ***	0.0944	0.8287	-2.848	0.0174
Travel agency	-1.5809 **	0.2058	0.7072	-2.235	0.0822
<b>Total fixed assets</b>					
Less than 3 million baht	-1.3950 ***	0.2479	0.2492	-5.597	0.0001
3-10 million baht	-0.5341 **	0.5862	0.1889	-2.827	0.0184
11-20 million baht	0.1468	1.1580	0.2059	0.713	0.7756
<b>Net income (Annual)</b>					
Less than 1 million baht	2.0970 ***	8.1390	0.6327	3.314	0.0041
1-5 million baht	1.1330 **	3.1040	0.5545	2.043	0.1240
6-10 million baht	0.6272	1.8720	0.5321	1.179	0.4992
<b>Business location (Region)</b>					
Northern region	0.1030	1.1085	0.3958	0.260	0.9667
Eastern region	-0.0641	0.9379	0.4825	-0.133	0.9912
Southern region	-0.8238 *	0.4388	0.4066	-2.026	0.1284
<b>Business premises</b>					
Rent/Lease a business premises	1.3030 *	3.6810	0.6136	2.124	0.1049
Own a business premises	1.8060 **	6.0860	0.5696	3.171	0.0066
Decreased revenue/income change (%)	0.0361 ***	1.0370	0.0103	3.493	0.0022
<i>Epidemic and business adaptation factors</i>					
Perceiving the risks of coronavirus impact (5 level)	0.7325 *	2.0803	0.3386	2.163	0.0963
<b>The current operation of a business</b>					
Open as usual and regular working hours	0.1358	1.1454	1.2465	0.109	0.9941
Open with shortened working hours and days	0.5955	1.8140	1.2522	0.476	0.8931
Not accepting customers but the employees are still working such as work from home and delivery.	0.7715	2.1630	1.4568	0.530	0.8692
<b>The business operation during the first wave</b>					
Open as usual and regular working hours	0.0361	1.0367	0.4863	0.074	0.9972
Open with shortened working hours and days	0.6362 *	1.8893	0.3150	2.020	0.1301
<b>Business model change from the coronavirus pandemic</b>					
Operate while adapting to social distancing	-0.5172 **	0.5962	0.1804	-2.866	0.0164
Move into new products and services that have high demand during the coronavirus pandemic	-0.1447	0.8653	0.1811	-0.799	0.7267
Operate through online markets or social media	-0.1280	0.8799	0.1832	-0.698	0.7834
Discussed with employees to reduce their salary to keep all employees	0.1338	1.1430	0.2134	0.627	0.8216

(continued)



Table 2. (continued)

Variable	Coefficient	Hazard ratio	SE	Z	MBF
<i>Employment factors</i>					
<b>Number of employees</b>					
Full-time employees	-0.0037	0.9963	0.0044	-0.838	0.7022
Part-time employees	0.0373 *	1.0380	0.0175	2.128	0.1032
<b>Labor Management</b>					
Reduced working hours to lower the unemployment rate	0.1129	1.1200	0.6731	0.168	0.9860
Doing everything as usual.	-1.3800 *	0.2516	0.7674	-1.798	0.1985
Lay off employees (Part/All)	-0.4284	0.6515	0.6384	-0.671	0.7984
<b>Dismissal of employees</b>					
Number of laid-off permanent employees.	0.0101	1.0101	0.0185	0.545	0.8615
Number of laid-off part time employees.	-0.0571 *	0.9445	0.0298	-1.913	0.1595
Number of employees expected to be laid off in the next 2 months.	0.0097	1.0097	0.0225	0.431	0.9113
<i>Target market factor</i>					
<b>Target customer</b>					
Domestic tourists' ratio	-0.3691 **	0.6913	0.1281	-2.881	0.0157
<b>Change in the number of tourists</b>					
Less than 25% reduction	1.6781	5.3555	1.2547	1.337	0.4089
Reduced by about 25-50%	2.2858 **	9.8340	1.0561	2.164	0.0961
Decreased by more than 50%	2.6165 **	13.6878	0.9768	2.679	0.0277
<i>Logistics and Supply Chain Factors</i>					
Import	-0.8983	0.4072	0.5523	-1.426	0.3314
Export	0.1855	1.2038	0.7406	0.251	0.9691
Import and export	-2.5682 **	0.0767	0.9537	-2.693	0.0266
<i>Government supporting policies factor</i>					
Low-interest loan measures, Soft Loan, 2% interest, and a 6-month moratorium on debt.	1.9262 ***	6.8636	0.5292	3.640	0.0013
Tax deductible measure for businesses that still employ all employees.	-0.3166	0.7286	0.4756	-0.666	0.8013
Measures to reduce contributions to the Social Security Fund for employees and employers	-1.3616 ***	0.2563	0.3937	-3.458	0.0025
Relaxation measure for personal and business tax filing	0.5853 .	1.7955	0.3338	1.453	0.3350
Measures to reduce withholding tax rates from services	-0.4701	0.6249	0.3526	-1.333	0.4112
Measures to extend the period of payment of income taxes (such as VAT) and excise taxes	-0.7115	0.4909	0.4460	-1.495	0.3601
Measures to stimulate domestic tourism by adding benefits for registrants "We travel together"	0.3892	1.4758	0.3881	1.003	0.6048
Not participating in any project	0.5341	1.7060	0.4749	1.125	0.5313

**Note:** The minimum Bayes factor (MBF) value provides the strength of evidence against the null hypothesis. It is the smallest possible Bayes factor for the point null hypothesis against the alternative within the specified class of alternatives [17]. \*\*\*, \*\*, and \* denote MBF, by 0.0001-0.01 MBF is decisive evidence, 0.01-0.1 is strong evidence, and 0.1-0.33 is moderate evidence rejecting the null hypothesis, respectively

## 4.2 Survival Path Analysis

This subsection presents an illustration of the survival probability over time. Following the previous parts, we employ the Kaplan-Meier estimator to evaluate

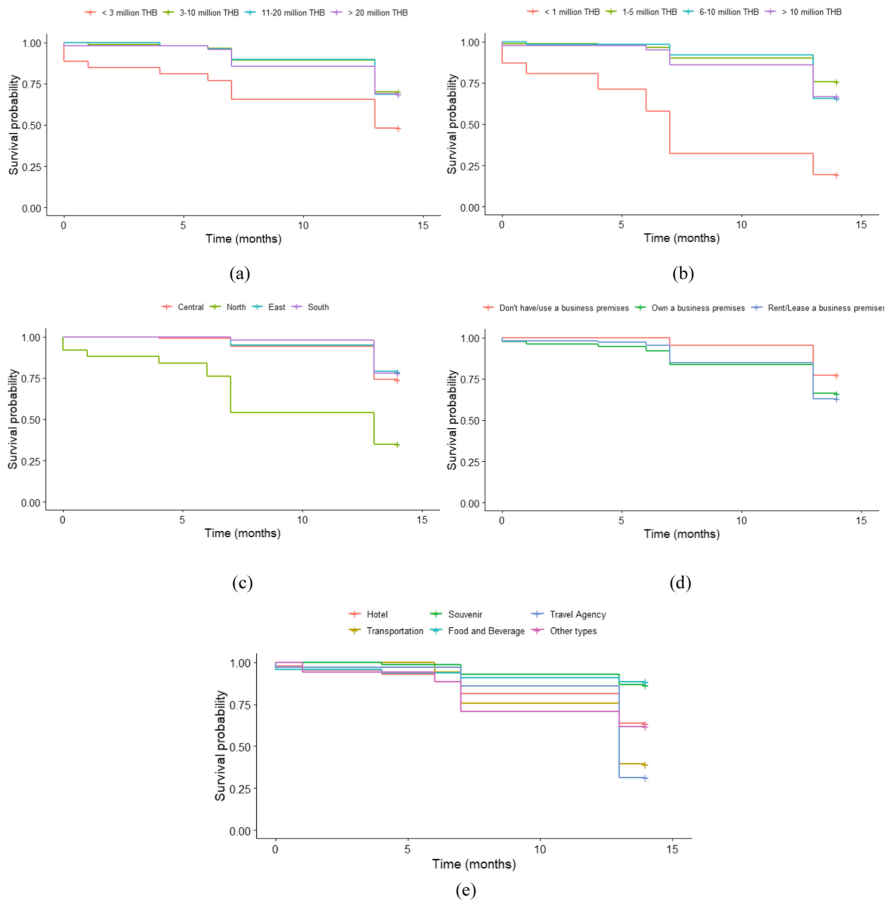
the impact of the various individual variables on the tourism business's survival. Only significant factors proposed by the Cox regression model are considered for the survival path analysis. We then plot the survival probability of the significant factor over the period as shown in Fig. 2.

The survival probability for each significant factor is illustrated in Figs. 2, 3, 4, 5 and 6. Figure 2 shows the survival curves for various types of tourism-serving establishments with varying total fixed assets, net income, locations, and types of business premises. It can be observed that the survival path in all variables has a stair-shaped downward slope in the first six months and a horizontal slope until the twelfth month. Moreover, we observe that the survival probability drops sharply after the sixth month, especially in total fixed assets of less than 3-million-baht businesses, a net income of less than 1 million baht businesses, businesses in the northern region, rent/lease business premises, and transportation and hotel business types. Then, the survival probability again sharply drops after the twelfth month, particularly in the travel agency business. These imply that the fundamental factors influence the survival probability of the tourism business in various categories only in the first six months. Thus, business adaptation and government support should be implemented during the first six months, particularly in the northern region. Furthermore, the risk of the crisis becomes stronger in businesses with low total fixed assets and worsens in low-income businesses indicating that small businesses are the ones most affected by the crisis.

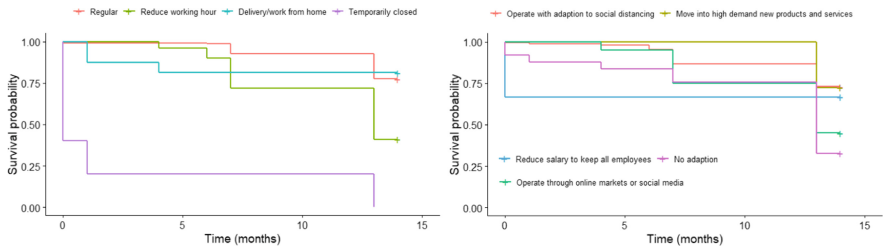
Figure 3 illustrates the survival probability of current business operations and their adaptation. The survival probability of a temporarily closed business is apparently lower than the other labor employment strategies. We also find that reducing working hours and keeping regular working hours are the best helpers for businesses to survive through the next 12 months. Meanwhile, businesses' adaptation to social distancing and switching to high-demand products and services give the highest survival probabilities at 78.2% and 72.4%, respectively.

The survival probabilities of the labor-management factors are depicted in Fig. 4. We discover that the temporary closure will cause the business not able to survive through the following year. Furthermore, reducing working hours can result in a dramatic decline of survival probability in month six and the subsequent stability through month 12, showing that employment adaptation should be taken within the first six months. Moreover, labor-management as usual should be continued if the business would like to survive beyond 12 months, considering the highest survival probability at 89.90%.

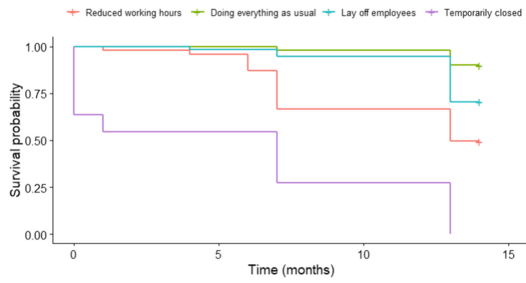
Figure 5 depicts the survival probabilities of various types of import/export of goods and services business from the Kaplan-Meier estimator. The overall result is in harmony with that obtained from the Cox regression model. Figure 5 shows that businesses having both exporting and importing components in their supply chain have the highest probability of surviving with a rate of 95.83%, implying that more distribution channels can increase the business survivability.



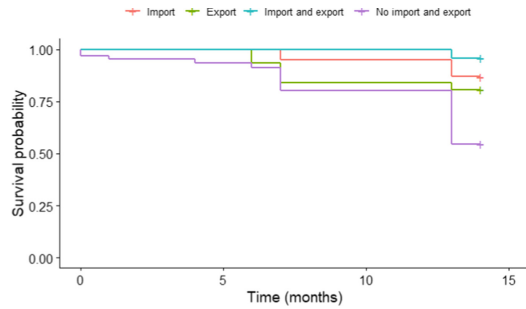
**Fig. 2.** Survival probability of tourism businesses in different total fixed assets(a), net income(b), regions(c), types of business premises (d), and business type (e)



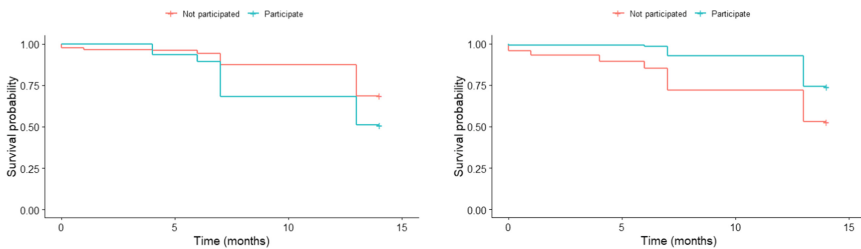
**Fig. 3.** Survival probability of tourism businesses in different current business operation (Left) and business adjustments (Right)



**Fig. 4.** Survival probability of tourism businesses in each labor management operation



**Fig. 5.** Survival probability of tourism businesses in different types of import/export of goods and services



**Fig. 6.** Probability of survival of tourism businesses with participation in debt-related (Soft loan) measure (Left) and Social Security Fund contributions reduction measure (Right)

Unexpectedly, Fig. 6 reveals that those businesses participating in the debt-relieving measures have a survival probability after 12 months (51.1%) lower than their non-participant counterpart, indicating that while these financial relief measures, such as Low-interest loans, Soft Loan, 2% interest, and a 6-month moratorium on debt, help businesses to avoid bankruptcy in the short term, they might not help them to survive in the long run. Meanwhile, participating in the

Social Security Fund contribution reduction scheme can increase the probability of business survivability to 74.00%.

## 5 Conclusion

This research regards the presence of the coronavirus crisis as the major cause of aggravating business failure in Thailand, especially in the tourism sector. Since lockdowns and travel restrictions have drastically reduced the number of tourists, the tourism-related business establishments that rely heavily on tourists have seen a substantial decline in revenue [18]. Undeniably, a strong possibility that many businesses might collapse in the upcoming months or years provides a terrific opportunity to assess the resilience factors behind these companies. Thus, the key challenge in this research is to investigate the role of factors supporting and inhibiting the probability of tourism businesses surviving during the coronavirus pandemic. Our study focuses on six influencing factors comprising firm's characteristics, ongoing crisis-related business adaption, target customer, supply chain, employment, and government support policies. All data are obtained by interviewing 400 tourism businesses in Thailand. For the research methodology, we start with identifying the key risk factors of non-survival businesses using the Cox proportional hazards model. Then, we use the Kaplan-Meier estimator to estimate the businesses' survival probability of individual significant parameters across survival periods.

First, our regression findings reveal that the Food and Beverage business type has the lowest risk of non-survival because their products are an essential part of daily life demand all the time. The fewer total fixed assets, the lesser of failure, indicating high business adaptability from having the low sunk cost. Businesses in the Southern region have a lower risk of non-survival than those in the Central. Undoubtedly, lower annual net income businesses appear to have a high probability of failure just like a decrease in revenue can worsen the business's survival probability. These results imply that having less liquidity might not be sufficient for the operation in a crisis. In business strategy factors, we find that businesses can accurately perceive the risk of failure. Therefore, this business risk self-awareness can be used as an indicator to monitor high-risk businesses for effective provision of government support. Additionally, a business model with social distancing can increase the possibility of survivability. For employment, an increase in the number of laid-off part-time employees and operation without working hours' reduction ameliorate the business's survival probability. In terms of target market factors, we discover that domestic tourist ratio growth and having both exporting and importing in the business supply chain can increase a firm's possibility of survival. Lastly, the debt-relieving measures worsen the probability of business survival, implying that these policies only slow down the bankruptcy but do not help the company to survive. On the contrary, the contribution reduction measure applied to the Social Security Fund can decisively increase the probability of survivability.

Second, the individual factors survival path analysis produced by the Kaplan-Meier estimator is consistent with the Cox regression results. Moreover, in terms

of the survival probability path over the year, all variables exhibit a stair-shaped downward slope in the first six months, followed by a dramatic decline, and then a horizontal stability through month 12, before a sharp drop. Figures 2-6 depict that all significant factors have a heterogeneous influence on the survival probability of the tourism business during the first six months and have a severe effect after the sixth month through twelfth. Thus, any business adaptation including government support should be promptly implemented within the first six months or at the latest within a year.

Lastly, this research data has certain limitations. We use the firm owner perspective to extrapolate business failure events. Hence, the survival probability might not accurately reflect the firm's insolvency. In other words, extrapolation and occurred results might not coincide. Therefore, using the actual data of business failure due to the coronavirus pandemic is intriguing for further study.

**Acknowledgments.** The authors are grateful to the Centre of Excellence in Econometrics, Chiang Mai University, and Faculty of Management Sciences, Chiang Mai Rajabhat University, for financial support. They are also grateful to Dr. Laxmi Worachai for her helpful comments and suggestions.

## References

1. United Nation World Tourism Organization [UNWTO]. International tourism highlights, 2020 edition. UNWTO (2020)
2. Tourism Authority of Thailand. Annual year report, 2021 edition. Thailand Ministry of Tourism and Sports (2021)
3. Leurcharusmee, S., Maneejuk, P., Yamaka, W.: A survival analysis of Thai micro and small-sized enterprises: does the COVID-19 pandemic matter? *J. Bus. Econ. Manage* **23**, 1211–1233 (2022)
4. Webster, A., Khorana, S., Pastore, F.: The labour market impact of COVID-19: early evidence for a sample of enterprises from Southern Europe. *Int. J. Manpower* **43**, 1054–1082 (2021)
5. Weaver, R.L.: The impact of COVID-19 on the social enterprise sector. *J. Soc. Entrepreneurship* **14**, 1–9 (2020)
6. Varzaru, A.A., Bocean, C.G., Cazacu, M.: Rethinking tourism industry in pandemic COVID-19 period. *Sustainability* **13**(12), 6956 (2021)
7. Pramana, S., Paramartha, D.Y., Ermawan, G.Y., Deli, N.F., Srimulyani, W.: Impact of COVID-19 pandemic on tourism in Indonesia. *Curr. Issues Tourism* **25**, 1–21 (2021)
8. Shafi, M., Liu, J., Ren, W.: Impact of COVID-19 pandemic on micro, small, and medium-sized Enterprises operating in Pakistan. *Res. Global.* **2**, 100018 (2020)
9. Huynh, D.V., Truong, T.T.K., Duong, L.H., Nguyen, N.T., Dao, G.V.H., Dao, C.N.: The COVID-19 pandemic and its impacts on tourism business in a developing city: insight from Vietnam. *Economies* **9**(4), 172 (2021)
10. Ma, Z., Liu, Y., Gao, Y.: Research on the impact of COVID-19 on Chinese small and medium-sized enterprises: evidence from Beijing. *PLoS ONE* **16**(12), e0257036 (2021)
11. Gregurec, I., Tomičić Furjan, M., Tomičić-Pupek, K.: The impact of COVID-19 on sustainable business models in SMEs. *Sustainability* **13**(3), 1098 (2021)

12. Cox, D.R.: Regression models and life-tables. In: Breakthroughs in Statistics, pp. 527-541. Springer, New York (1992). [https://doi.org/10.1007/978-1-4612-4380-9\\_37](https://doi.org/10.1007/978-1-4612-4380-9_37)
13. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958)
14. Walters, S.J.: What is a Cox model? From University of Oxford Clinical School Information Management Services Unit (2008). [www.jr2.ox.ac.uk/bandolier/painres/download/whatis/COX\\_MODEL.pdf](http://www.jr2.ox.ac.uk/bandolier/painres/download/whatis/COX_MODEL.pdf)
15. Gemar, G., Moniche, L., Morales, A.J.: Survival analysis of the Spanish hotel industry. *Tour. Manage.* **54**, 428–438 (2016)
16. Kavkler, A., et al.: Cox regression models for unemployment duration in Romania, Austria, Slovenia, Croatia, and Macedonia. *Rom. J. Econ. Forecast.* **10**(2), 81–104 (2009)
17. Maneejuk, P., Yamaka, W.: Significance test for linear regression: how to test without P-values? *J. Appl. Stat.* **48**(5), 827–845 (2021)
18. Office of SMEs Promotion. MSMEs and the COVID-19 overview, 2020 edition. OSMPE (2020)



# Impacts of Capital Structure on Microfinance Institutions' Risk: Evidence from Low- and Middle-Income Countries

Thuy T. Dang<sup>1</sup>, Nguyen Tran Xuan Linh<sup>2</sup>, Hau Trung Nguyen<sup>3</sup>(✉),  
and Dinh Cong Hoang<sup>4</sup>

<sup>1</sup> Institute for Indian and Southwest Asian Studies, Vietnam Academy of Social Sciences, 176 Thai Ha Street, Dong Da, Hanoi, Vietnam

thuy0183@gmail.com

<sup>2</sup> University of Finance – Marketing, 778 Nguyen Kiem Street, Phu Nhuan, Ho Chi Minh, Vietnam

ntxlinh@ufm.edu.vn

<sup>3</sup> Banking Strategy Institute, State Bank of Vietnam, 504 Xa Dan Street, Dong Da, Hanoi, Vietnam

hau.nguyentrung@sbv.gov.vn

<sup>4</sup> Institute for Africa and Middle East Studies, Vietnam Academy of Social Sciences, 176 Thai Ha Street, Dong Da, Hanoi, Vietnam

hoang0108@gmail.com

**Abstract.** This study aims to evaluate the impact of capital structure and other factors on risk of microfinance institutions (MFIs) in 26 low-middle-income countries during 2014–2017. Empirical results show that MFIs with high financial leverage tended to face with higher risks. Thus, authorities in such countries should increase supervision of MFIs with modest equity in order to ensure the stable and sustainable operation of the MFIs system. In addition, the research also recognizes that size of MFIs, number of borrowers, total assets and debt-to-equity ratio had a negative impact on MFIs control of bad debts. Female borrowers, borrowers in rural areas and number of branches tended to lessen MFIs bad debts. Lending to female customers can both bring benefits to MFIs and empower women in society.

**Keywords:** microfinance institutions · capital structure · credit risk · low- and middle-income countries

## 1 Introduction

Sustainable economic growth is always a strategic goal of all countries in the world (Osborn et al., 2015). To achieve this goal, all sectors of the economy must receive support from the national financial system. Thus, governments pay much attention to financial inclusion for every people in every region, especially low-income and disadvantaged groups. Over past decades, developing countries have witnessed the emergence

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 654–666, 2024.

[https://doi.org/10.1007/978-3-031-35763-3\\_46](https://doi.org/10.1007/978-3-031-35763-3_46)



and strong emergence of MFIs. Robinson (2001) explains that a MFI is one kind of financial institution that specializes in providing small-scale financial services to business owners in urban and rural communities. Organization for Economic Co-operation and Development (OECD, 2013) regards MFIs as a pillar for financial inclusion, and Okumadewa (1999) believes that the development of MFIs will ensure the efficiency of poverty reduction projects and programs.

For low-middle-income countries, the important goal is socio-economic development, poverty reduction and sustainable economic growth. The strength of MFIs plays a key role to achieve this goal because they are an effective tool in poverty reduction, narrowing the supply-demand gap in accessing necessary financial services of small and micro enterprises in low-middle-income countries (Karam, 2014). Unlike traditional credit institutions, MFIs receive money and support from government to provide non-credit services such as capacity building, vocational training, product marketing, etc. to their customers to improve effective usage of loans (Hartungi, 2007). These features ensure the suitability of MFIs for low-income and inadequate education customers in society.

Because of its important role in poverty and inequality reduction, ensuring the sustainable operation of MFIs has received much attention from both governments and researchers. For this reason, the authors will analyze the impact of capital structure on the risk of MFIs and propose recommendations to ensure sustainable operation of MFIs in low-middle-income countries.

## 2 Literature Review

Basic functions of capital in the financial sector are as follows: (i) it is a buffer to absorb losses, (ii) it helps to increase the confidence of depositors, (iii) it represents the level of risk borne by the owners, (iv) it indicates the minimum cost financing method used by the bank (Ayaydin and Karakaya, 2014). Basically, capital structure is the combination of equity and debt in total capital of a company.

Different economic theories provided different forecasts about capital and risk of credit institutions. Among these theories, “regulatory hypothesis” implies that credit institutions with modest equity capital tend to face with requirement to increase capital by decrease dividends to shareholders. Anginer and Demirgüç-Kunt (2014) explain that credit institutions with high capital ratio can cope with income shocks and ensure financial ability to settle down deposit withdrawals and other agreements of customers. They also explain that higher capital buffers help credit institution to be more prudent and wiser in their investment decisions. Accordingly, the policy of “more skin in the game” helps to improve the monitoring and screening of credit institution risks. Higher capital ratio will decrease the pressure on the liabilities of credit institutions and the risk of having to call for government bailouts (Beck et al., 2013). This view is supported by a lot of empirical evidences. Jacques and Nigro (1997) demonstrated that high risk-based capital measures can decrease the risk of credit institutions. Similarly, Aggarwal and Jacques (1998) used Federal Deposit Insurance Corporation (FDIC) data for credit institutions from 1990 to 1993 and showed that credit institutions tend to maintain capital-to-required reserves ratio as a way to prevent breakdowns from unexpected dire situations. Editz et al (1998)

analyzed the relationship between regulations and stability of credit institutions, using research data from British credit institutions. They showed that reserve regulations have positive impact on the stability and soundness of credit institution system and it do not distort the lending ability of commercial credit institutions.

“Moral hazard” is another theory that replaces “regulatory hypothesis”. It believes that unqualified credit institutions tend to take excessive risks in order to maximize their shares value at the expense of depositors. A similar framework to analyse bank capital and risk is the too-big-to-fail hypothesis that has generated moral hazard behavior leading to excessively risky activities under deposit insurance and government bailouts. Koehn and Santomero (1980) argued that a higher capital ratio will increase volatility of total risk of credit institution sector. Blum (1999) uses dynamic framework to explain that if raising capital to meet future standards is too costly, the only solution for credit institutions is to increase the risk of their portfolio with the expectation of present high return will be able to meet the minimum capital requirements in the future.

Studies on relationship between capital structure and risk of credit institutions are mostly done with commercial banks instead of MFIs and have inconsistent results.

Based on regulatory hypothesis framework, Berger and Bouwman (2013) found a positive effect of equity on the survivability of US small banks during 1981:Q1 - 2010:Q4 crises. With the same framework, Tan and Floros (2013) analysed Chinese commercial banks data, and Anginer and Demirgüç-Kunt (2014) analysed commercial banks data from 48 countries and found a negative relationship between equity ratio and bank risk.

In contrast, based on “moral hazard” theory and too-big-to-fail hypothesis, Keeton and Morris (1987) affirmed that bank capitalization played an important role in determining the loan risk level. With data from US banks in the period of 1979–1985, Keeton and Morris (1987) proved that credit institutions with high equity-to-asset ratio tend to have higher bad debt level. Similarly, Iannotta et al. (2007) found positive correlation between capital and loan losses when analyzing the relationship between capital size and risk of large European banks from 1999 to 2004.

For MFIs, Morduch (1999) claimed that group lending may reduce risk by reducing adverse selection and moral hazard. Group loan agreements effectively bind loan co-signers, minimize information asymmetry problems between lenders and borrowers. Co-signers are encouraged to monitor each other, exclude risky borrowers, increase repayment rates. Similarly, analyzing 148 MFIs during 2001 to 2006, Wagner and Winkler (2013) found that group lending contracts significantly improved portfolio quality. However, group lending causes MFIs to make a trade-off between risk and performance efficiency because group lending may increase MFIs’ management costs.

With a sample of 37 MFIs in the period of 2001–2003, Crabb and Keller (2006) examined main risk factors of loan portfolio, including size of MFIs and macroeconomic factors. They found that group lending method may reduce loan portfolio risk. They also believed that lending to women can reduce MFIs’ portfolio risk. Research by Saravia-Matus and Saravia-Matus (2015) confirmed that repayment performance of women was better than that of men in a sample of Nicaragua MFIs.

Ayayi (2012) studied the determinants of credit risk for MFIs in Vietnam, East Asia and the Pacific. Research shows that liquidity will limit the credit risk of MFIs while the

size of the total loan portfolio and inefficiency have a negative impact on the credit risk of MFIs in this area.

Based on survey data from 250 borrowers from MFIs in Ghana, Addae-Korankye (2014) confirmed that the factors leading to high NPLs at microfinance institutions are high-interest rates, poor appraisal, inappropriate loan sizes, and improper client selection.

Based on data from borrowers at MFIs in Ghana from 2011 to 2014, Mensah (2016) pointed out factors that increase bad debt risk at MFIs, including low education rates, ineffective monitoring, poor loan appraisal, and crop failure.

Based on a dataset of 607 microfinance institutions in 87 developing countries, Zamore et al. (2021) found that high operating costs force MFIs to raise interest rates, which has hurt the ability of customers to pay debts. In addition, the study also found that group lending helps reduce operating costs, thereby reducing the risk of increased bad debt. Besides, MFIs that focus on lending in rural areas will be more effective than serving both urban and rural clients and thereby help reduce bad debts. The large size of the MFIs also increases the operating costs of the microfinance institutions and thus increases the risk of bad debt of them.

In this study, we analyze the impact of capital structure on risk of MFIs in low- and middle-income countries.

### 3 Research Model and Research Method

The authors will analyze the impact of capital structure on the risk of MFIs in low- and middle-income countries through testing two hypotheses.

Hypothesis 1: Debt to equity ratio (financial leverage) increases MFIs risk.

Hypothesis 2: Debt to equity ratio reduces MFIs risk.

Although there are many indicators of risk measurement of MFIs, due to limitation of research data, in this study, the authors use bad debt ratio as a measurement of MFIs risk. Moreover, because of available data in IMF survey in 59 low-middle-income countries, we only consider 26 low-middle-income countries in annual report of Global Outreach and Financial Performance Standards (Global Outreach and Financial Performance) survey conducted by MixMarket in the period of 2014 to 2017. Besides capital structure, which is measured by debt-to-equity ratio, we also analyze other factors:

- *Female borrowers*: MFIs main customers are low-income, under-educated and disadvantaged people in society, especially women. Studies of Crabb and Keller (2006), Saravia-Matus and Saravia-Matus (2015) examined the impact of loans for women on MFIs' performance and showed that targeting female borrowers is of great significance in empowering and enhancing women's role in the society.
- *Borrowers in rural areas*: Most borrowers in rural areas are disadvantage people, small borrowers, thus, they are difficult to access bank credit. According to Quayes (2015), people in rural areas of low-middle-income countries and transition economies like Vietnam and India are facing difficulties in land scarcity for farming and they tend to shift to small business. They are considered the target customer group of MFIs.

- *Total assets of MFIs*: Total assets represent the strength of MFIs and their ability to deal with information asymmetry, leading to a lower bad debt level. Smaller MFIs have fewer resources for efficient credit analysis. Furthermore, the size of MFIs can be an indicator of increasing diversification opportunities, so that MFIs' bad debts will decrease.
- *Number of branches*: MFIs with large branches are easier to collect customer information, reduce information asymmetry, provide further support to customers in doing business. This fact may improve financial capacity of MFIs' borrowers.
- *Number of borrowers*: with a plenty of small-value loans, MFIs' credit officers are in charge of many loans at the same time which tends to have negative impacts on supervisory advisory ability of loan officers.
- *Economic growth*: Abuzayed et al. (2018) showed that economic growth will increase the disposable income of individuals and households, which improve the borrower's ability to fulfill financial obligations. Moreover, with an increase in income, people also tend to consume more, leading to a positive impact on companies' performance results, thereby improving their financial capacity. Thus, economic growth will reduce the bad debt ratio (Kjosevski et al., 2019).
- *Inflation reduces real income of individuals and households*. Their reduction in consumption as inflation reaches high level may have negative impacts on companies' profit which indirectly caused and increase in bank's bad debt (Kastrati, 2011; Abuzayed et al., 2018).

The research model has the form:

$$\text{PAR30} = \alpha_1 + \alpha_2 \text{LEV} + \alpha_3 \text{ASS} + \alpha_4 \text{FEM} + \alpha_5 \text{RUR} \\ + \alpha_6 \text{OFI} + \alpha_7 \text{BOR} + \alpha_8 \text{GDP} + \alpha_9 \text{INF} + \varepsilon$$

$$\text{PAR90} = \beta_1 + \beta_2 \text{LEV} + \beta_3 \text{ASS} + \beta_4 \text{FEM} + \beta_5 \text{RUR} \\ + \beta_6 \text{OFI} + \beta_7 \text{BOR} + \beta_8 \text{GDP} + \beta_9 \text{INF} + \varepsilon$$

Since performance indicators is calculated as a percentage, average loan value and number of borrowers will be in form of natural logarithms to calibrate large skewed economic variables for statistical analysis.

Most prior researches used frequency approach, a prior information is not available. However, 2057 observations in the sample is large enough; hence prior information does not affect the posterior distribution too much. In this case, Block et al. (2011) proposed Gaussian standard distribution with different prior information (simulation of a priori information) and carried out Bayesian factors to choose a simulation with best a prior information.

The simulations in Table 2 show decreasing levels of a prior information. Simulation 1 has strongest a prior information and Simulation 5 has weakest a prior information.

Similar to model 2 with dependent variable PAR90, we also build 5 simulations (from simulation 6 to simulation 10) among wick simulation 6 has strongest a prior information ( $\beta_i \sim N(0,1)$ ) and simulation10 has weakest a prior information ( $\beta_i \sim N(0,10000)$ ).

**Table 1.** Variable description

Research variable		Symbol
Dependent variable	Debt overdue over 90 days	PAR90
	Debts overdue over 30 days	PAR30
Independent variable	Financial leverage	LEV
	Total assets of MFIs	ASS
	Women borrowers	FEM
	Loan customers in rural areas	RUR
	Number of branches of microfinance institution	OFI
	Number of customers borrowing	BOR
	GDP growth rate	GDP
	Inflation Rate	INF

Source: The authors

**Table 2.** Simulation summary

Likelihood	$PAR30 \sim N(\mu, \sigma)$
Prior distributions:	
Simulation 1	$\alpha \sim N(0, 1)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 2	$\alpha \sim N(0, 10)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 3	$\alpha \sim N(0, 100)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 4	$\alpha \sim N(0, 1000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$
Simulation 5	$\alpha \sim N(0, 10000)$ $\sigma^2 \sim Invgamma(0.01, 0.01)$

Source: The authors

In next step, the authors carried out Bayesian regression for 10 simulations, then performed Bayesian factor analysis and Bayesian test model. These techniques are proposed by StataCorp LLC (2019) to select the simulation with the best a prior information. Basically, the Bayesian factor provide a tool to compare probability of a particular hypothesis (a prior information) to probability of another hypothesis. It is a strength measure of an evidence in favor of one theory among competing (information a prior) theories. Bayesian analysis provides average Log BF, Log ML and average DIC (Deviance Information Criterion). Posterior Bayesian compares posterior probability of simulations with different

a prior information. Based on research data and proposed a prior information, the authors will choose simulation with greatest posterior probability  $P(M|y)$ . In short, the authors build 5 simulations with 5 different priori information. Bayesian factor analysis and posterior Bayesian test helps us to choose a simulation with most suitable prior information. The simulation with largest Log BF, Log ML average, minimum DIC mean and the largest  $P(M|y)$  will be selected.

**Table 3.** A Bayesian factor test and a model test

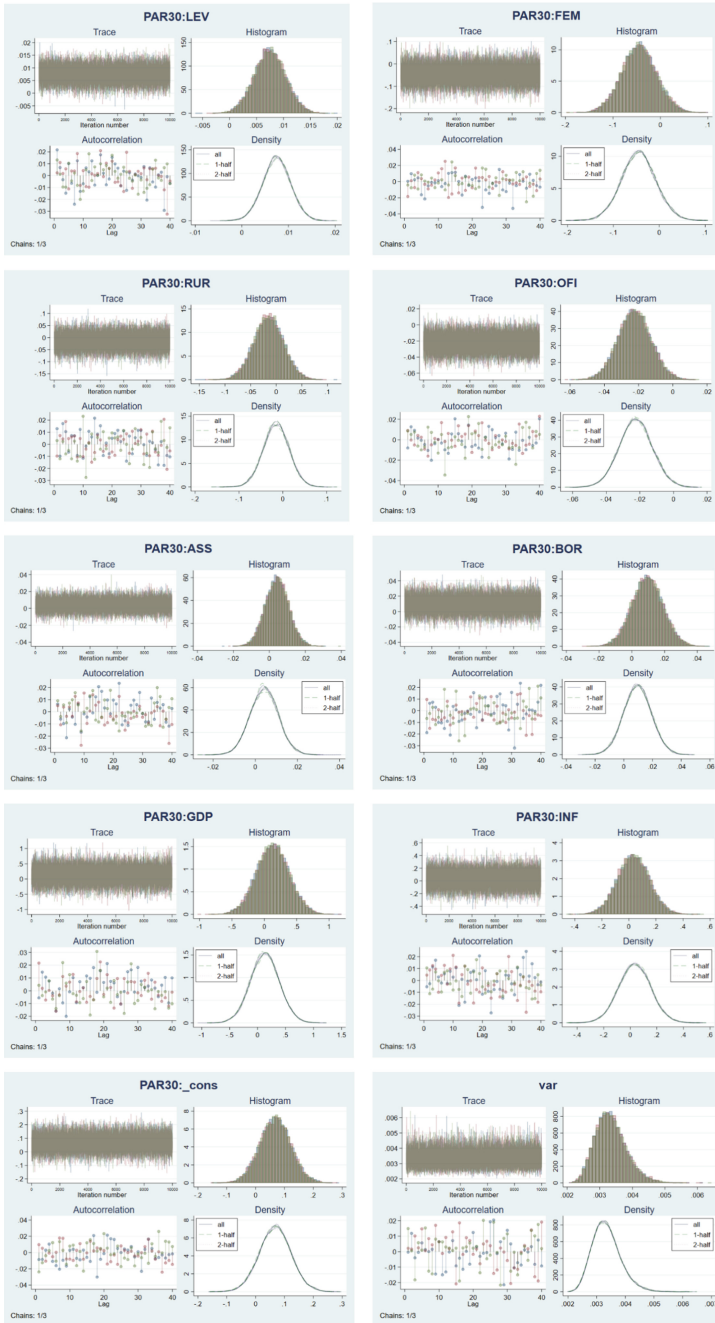
Model 1					
	Chains	Avg DIC	Avg log(ML)	log(BF)	$P(M y)$
simulation1	3	-295.339	111.426		1.000
simulation2	3	-295.063	101.150	-10.276	0.000
simulation3	3	-295.009	90.787	-20.639	0.000
simulation4	3	-295.200	80.366	-31.060	0.000
simulation5	3	-295.065	69.947	-41.479	0.000
Model 2					
	Chains	Avg DIC	Avg log(ML)	log(BF)	$P(M y)$
simulation6	3	-382.9462	148.1828		1.000
simulation7	3	-382.7491	137.9575	-10.2253	0.000
simulation8	3	-382.8882	127.5187	-20.6641	0.000
simulation9	3	-382.7656	117.1876	-30.9952	0.000
simulation10	3	-382.7412	106.8183	-41.3645	0.000

Source: Authors' calculation

Table 3 shows that simulation 1 is the most suitable prior information simulation. Results of posterior test also show that simulation 1 has superiority over other simulations. Thus, simulation 1 with prior information  $N(0, 1)$  would be selected. Similarly, simulation 6 with a priori  $N(0, 1)$  is the best fit for model 2.

Bayesian analysis is simulated through the Markov Chain Monte Carlo (MCMC), therefore, to ensure stability of Bayesian regression, MCMC series must converge, which means MCMC series must ensure stationarity. StataCorp LLC (2019) proposes that MCMC series convergence test can be conducted through convergence diagnostic graph.

According to StataCorp LLC (2019), MCMC series convergence diagnostic graph includes trace plot, histogram, autocorrelation, density plot. Trace plot helps to track historical display of parameter value via repetitions of the series. Figure 1 shows that trace plot fluctuates around mean value, so MCMC series is stationary because convergence condition is satisfied. Besides, autocorrelation chart fluctuates around a level below 0.02, showing the agreement with density distribution simulation and lagging within effective limit (StataCorp LLC, 2019). According to StataCorp LLC (2019), posterior distribution plot and density estimate show normal distribution of parameters. Normal histogram shape shows that Bayesian regression ensure stability. Thus, MCMC series



**Fig. 1.** Convergence diagnostic graph Source: Authors' calculation

meets the convergence condition from Fig. 1. For model 2, the authors also obtained similar results.

## 4 Discussion

In addition to graphical convergence diagnostics, StataCorp LLC (2019) also recommends testing via Average Acceptance Rate; Average minimum efficiency; and maximum Gelman-Rubin  $R_c$ . Table 4 shows that model's acceptance rate reaches 1, model's minimum efficiency is 0.94, much greater than allowable level of 0.01. Maximum  $R_c$  value of coefficients is 1, showing that MCMC chains in Table 4 satisfy the convergence requirements (Gelman and Rubin, 1992). We also consider Monte-Carlo Standard Error (MCSE) criterion to test the stability of MCMC chains, results in Table 4 show that the MCSE value meets the optimal level (Flegal et al., 2008).<sup>1</sup>

Regression results show that financial leverage (LEV) tends to increase the risk of MFIs when it has a positive effect on both PAR30 and PAR90. Other factors including total asset size (ASS) and number of borrowers (BOR) also tend to increase PAR30 and PAR90, while female borrowers (FEM), customers loans in rural areas (RUR) and the number of branches of MFIs (OFI) help to reduce the risks of MFIs. For macro factors, GDP tends to increase the risk of MFIs as it increases both PAR30 and PAR90, while inflation (IMF) tends to increase PAR30 but decrease PAR90, so they We need to analyze this factor more closely. Unlike the frequency method, we can only determine the sign of the regression coefficient, but we cannot estimate the degree of certainty for this effect, whereas with the Bayes method we can calculate the probability of these impact trends occurring. This is considered to be the outstanding advantage of Bayesian method over frequency method.

Results in Table 5 show that financial leverage (LEV) ratio increases bad debt ratio PAR30 and PAR90 of MFIs with a probability of exceeding 99%. This is consistent with theoretical framework "risk management" and "moral risk" presented in Sect. 2. Bad debts of credit institutions tend to be higher in institutions with low equity ratio, similar to Keeton and Morris (1987), Berger and DeYoung (1997), Louzis et al (2010), Stolz and Wedow (2011) researches. This affirms the important role of owner's equity of MFIs in stabilizing the operation of MFI system.

Research results also show that female borrowers (FEM) tend to have a positive impact on risk control of MFIs with the probability of 89% for PAR30 and 97% for PAR90, similar to study of Saravia-Matus (2015). Bad debt ratio of borrowers in rural areas (RUR) is also lower than that in urban areas, but probability is only 69.7% for PAR30 and 68.4% for PAR90, showing this effect is not clear. Number of MFIs branches also helps to significantly improve bad debt control with a probability of nearly 99% for PAR30 and PAR90. Obviously, if MFIs network increases, their ability to access and screening borrowers is better and bad debt risk will be reduced. However, an increase in number of branches also lead to an increase in management cost.

<sup>1</sup> Flegal et al. (2008): the nearer MCSE approaches zero, the better; the stronger the MCMC chain, the better MCSE is; MCSE is acceptable if less than 6.5% standard deviation, and is optimal if less than 5% standard deviation.



**Table 4.** Bayesian simulation outcomes

	Mean	Std. Dev.	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
<b>PAR30</b>						
LEV	0.007	0.003	0.000	0.007	0.002	0.013
FEM	-0.045	0.037	0.000	-0.045	-0.118	0.029
RUR	-0.015	0.030	0.000	-0.015	-0.074	0.043
OFI	-0.022	0.010	0.000	-0.022	-0.041	-0.003
ASS	0.004	0.007	0.000	0.004	-0.009	0.017
BOR	0.009	0.010	0.000	0.009	-0.009	0.028
GDP	0.132	0.257	0.001	0.132	-0.368	0.636
INF	0.033	0.121	0.001	0.034	-0.205	0.270
_cons	0.069	0.055	0.000	0.069	-0.039	0.177
var	0.003	0.000	0.000	0.003	0.003	0.004
Avg acceptance rate	1.000					
Avg efficiency: min	0.981					
Max Gelman-Rubin Rc	1					
<b>PAR90</b>						
LEV	0.005	0.002	0.000	0.005	0.001	0.009
FEM	-0.048	0.025	0.000	-0.048	-0.098	0.002
RUR	-0.009	0.020	0.000	-0.010	-0.049	0.030
OFI	-0.015	0.007	0.000	-0.015	-0.028	-0.002
ASS	0.001	0.005	0.000	0.001	-0.008	0.010
BOR	0.007	0.007	0.000	0.007	-0.006	0.020
GDP	0.171	0.178	0.001	0.170	-0.180	0.526
INF	-0.067	0.081	0.000	-0.067	-0.228	0.092
_cons	0.058	0.037	0.000	0.058	-0.015	0.131
var	0.002	0.000	0.000	0.002	0.001	0.002
Avg acceptance rate	1.000					
Avg efficiency: min	0.979					
Max Gelman-Rubin Rc	1					

Source: Authors' calculation

Number of borrowers tend to reduce MFIs risks with probability of positive impact on PAR30 and PAR90 is 83.8% and 87.2%, respectively. This can be explained that when MFIs boost lending, increase number of borrowers, they have to lower borrowing conditions, so an increase in bad debt is inevitable. Table 6 also shows that large asset

**Table 5.** Posterior probability

	Mean	Std. Dev	MCSE
PAR 30			
Probability {PAR30:LEV} > 0	0.994	0.078	0.000
Probability {PAR30:FEM} < 0	0.886	0.318	0.002
Probability {PAR30:RUR} < 0	0.697	0.460	0.003
Probability {PAR30:OFI} < 0	0.988	0.108	0.001
Probability {PAR30:ASS} > 0	0.743	0.437	0.003
Probability {PAR30:BOR} > 0	0.838	0.369	0.002
Probability {PAR30:GDP} > 0	0.695	0.460	0.003
Probability {PAR30:INF} > 0	0.612	0.487	0.003
PAR 90			
Probability {PAR90:LEV} > 0	0.993	0.081	0.000
Probability {PAR90:FEM} < 0	0.971	0.167	0.001
Probability {PAR90:RUR} < 0	0.684	0.465	0.003
Probability {PAR90:ASS} > 0	0.592	0.491	0.003
Probability {PAR90:OFI} < 0	0.986	0.117	0.001
Probability {PAR90:BOR} > 0	0.872	0.334	0.002
Probability {PAR90:GDP} > 0	0.731	0.375	0.002
Probability {PAR90:INF} < 0	0.698	0.401	0.002

Source: Authors' calculation

size of MFIs also increases bad debt risk, however, the probability of this effect is low, only 74.3% for PAR30 and 59.2% for PAR90.

Impact of macro factors on MFIs risk is unclear when probability of positive impact of GDP on PAR30 is 69.5% and PAR90 is 73.1%. Meanwhile, positive impact of inflation (INF) on PAR30 is 61.1%, and its negative impact on PAR90 is 69.8%.

## 5 Conclusion

This research aims to assess impact of capital structure on MFIs credit risk in low- and middle-income countries. Research results show that MFIs with high financial leverage tend to have higher bad debts. MFIs with adequate source of capital can be active in business activities, increase ability to mobilize deposit, expand credit, and reduce financial burden. MFIs with low owner's equity level have to face with loan portfolio risk increases due to poor diversification loan portfolio. This implies that increasing equity capital size is a key factor in maintaining MFIs stability. In addition, authorities in these countries should increase supervision on MFIs with low equity or ask them to increase equity.

MFIs size, number of borrowers, total assets and debt-to-equity ratio have negative impact on credit risk control of MFIs, whereas effect of total assets is unclear because its probability is relatively low. Meanwhile, female borrowers, borrowers in rural areas and number of branches have a positive impact on reducing MFIs bad debts. However, increase in number of branches will increase operating costs and thereby reduce the operational efficiency of these organizations. Therefore, MFIs should not expand their branch operations aggressively. Loans to customers in rural areas have a positive impact on controlling MFIs bad debts, therefore, the authors suggest that MFIs in low- and middle- income countries should focus on rural market segment instead of spreading its scope of activities.

Finally, loans to female customers have a positive effect on reducing MFIs bad debt, therefore, MFIs should expand their lending to this target group. It may benefit MFIs and the whole society via women empowerment to affirm their role in low- and middle-income countries.

## References

- Abuzayed, B., Al-Fayoumi, N., Molyneux, P.: Diversification and bank stability in the GCC. *J. Int. Financ. Mark. Inst. Money* **57**, 17–43 (2018)
- Addae-Korankye, A.: Causes and control of loan default/delinquency in microfinance institutions in Ghana. *Am. Int. J. Contemp. Res.* **4**(12), 36–45 (2014)
- Aggarwal, R.K., Jacques, K.: Assessing the impact of prompt corrective action on bank capital and risk. *Econ. Policy Rev.* **4**(3), 23–32 (1998)
- Anginer, D., Demirgüç-Kunt, A., Zhu, M.: How does bank competition affect systemic stability? *J. Financ. Intermed.* **23**(1), 1–26 (2014)
- Ayayi, A.G.: Credit risk assessment in the microfinance industry. *Econ. Transit.* **20**(1), 37–72 (2012)
- Ayaydin, H., Karakaya, A.: The effect of bank capital on profitability and risk in Turkish banking. *Int. J. Bus. Soc. Sci.* **5**(1) (2014)
- Beck, T., Demirgüç-Kunt, A., Merrouche, O.: Islamic vs. conventional banking: business model, efficiency and stability. *J. Bank. Financ.* **37**(2), 433–447 (2013)
- Berger, A.N., DeYoung, R.: Problem loans and cost efficiency in commercial banks. *J. Bank. Financ.* **21**(6), 849–870 (1997)
- Berger, A.N., Bouwman, C.H.: How does capital affect bank performance during financial crises? *J. Financ. Econ.* **109**(1), 146–176 (2013)
- Blum, J.: Do capital adequacy requirements reduce risks in banking? *J. Bank. Financ.* **23**(5), 755–771 (1999)
- Block, J.H., Peter, J., Miller, D.: Ownership versus management effects on performance in family and founder companies: a Bayesian reconciliation. *J. Fam. Bus. Strategy* **2**, 232–245 (2011)
- Crabb, P.R., Keller, T.: A test of portfolio risk in microfinance institutions. *Faith Econ.* **47**(48), 25–39 (2006)
- Editz, T., Michael, I., Perraudin, W.: The impact of capital requirements on U.K. bank behaviour. *Reserve Bank New York Policy Rev.* **4**(3), 15–22 (1998)
- Flegal, J.M., Haran, M., Jones, G.L.: Markov chain Monte Carlo: can we trust the third significant figure? *Stat. Sci.* **23**(2), 250–260 (2008)
- Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**(4), 457–472 (1992)

- Iannotta, G., Nocera, G., Sironi, A.: Ownership structure, risk and performance in the European banking industry. *J. Bank. Financ.* **31**(7), 2127–2149 (2007)
- Jacques, K., Nigro, P.: Risk-based capital, portfolio risk, and bank capital: a simultaneous equations approach. *J. Econ. Bus.* **49**(6), 533–547 (1997)
- Hartungi, R.: Understanding the success factors of micro-finance institution in a developing country. *Int. J. Soc. Econ.* **34**, 388–401 (2007)
- Kastrati, A.: The Determinants of Non-Performing Loans in Transition Countries. Financial Stability Report, June, Central Bank of the Republic of Kosovo (2011)
- Narwal, K.P.: Impact of characteristics on outreach and profitability of microfinance institutions in India. *Int. J. Financ. Manag.* **4**(3), 50–57 (2014)
- Keeton, W.R., Morris, C.: Why do banks' loan losses differ? *Fed. Reserve Bank Kansas City Econ. Rev.* **72**(5), 3–21 (1987)
- Kjosevski, J., Petkovski, M., Naumovska, E.: Bank-specific and macroeconomic determinants of non-performing loans in the Republic of Macedonia: comparative analysis of enterprise and household NPLs. *Economic Research-Ekonomska Istraživanja* **32**(1), 1185–1203 (2019)
- Koehn, M., Santomero, A.: Regulation of bank capital and portfolio risk. *J. Financ.* **35**(5), 1235–1245 (1980)
- Louzis, D., Vouldis, A., Metaxas, V.: Macroeconomic and bank-specific determinants of non-performing loans in Greece: a comparative study of mort-gage, business and consumer loan portfolios. *J. Bank. Financ.* **36**(4), 1012–1027 (2010)
- Mensah, E.: Analysis of non-performing loans: a case study of Dunkwa area Teachers Co-Operative Credit Union (DATCCU) (Doctoral dissertation, Kwame Nkrumah University of Science and Technology) (2016)
- Morduch, J.: The promise of microfinance. *J. Econ. Lit.* **37**(4), 1569–1614 (1999)
- OECD (2013). Promoting Financial Inclusion through Financial Education: OECD/INFE Evidence, Policies and Practice. OECD Working Papers on Finance, Insurance and Private Pensions No. 34
- Okumadewa, F.: International agencies response to poverty situation in Nigeria. *Cent. Bank Niger. Bullion* **23**(4), 66–70 (1999)
- Osborn, D., Cutter, A., Ullah, F.: Universal sustainable development goals - Understanding the transformational challenge for developed countries. Report of a study by stakeholder forum (2015)
- Quayes, S.: Outreach and performance of microfinance institutions: a panel analysis. *Appl. Econ.* **2015** (2015)
- Robinson, M.S.: The Microfinance Revolution Sustainable Finance for the Poor, World bank Book (2001)
- Saravia-Matus, S.L., Saravia-Matus, J.A.: Gender issues in microfinance and repayment performance: the case of a Nicaraguan microfinance institution. *Encuentro: Revista Académica de la Universidad Centroamericana* **91**, 7–31 (2015)
- StataCorp LLC. (2019). Stata Bayesian Analysis Reference Manual Release 15. Statistical Software. Texas: College Station
- Stolz, S., Wedow, M.: Banks' regulatory capital buffer and the business cycle: evidence for Germany. *J. Financ. Stab.* **7**(2), 98–110 (2011)
- Tan, Y., Floros, F.: Risk, capital and efficiency in Chinese banking. *Int. Financ. Mark. Inst. Money* **26**, 378–393 (2013)
- Thach, N.N.: How to explain when the ES is lower than one? A Bayesian nonlinear mixed-effects approach. *J. Risk Financ. Manag.* **13**, 21 (2020)
- Wagner, C., Winkler, A.: The vulnerability of microfinance to financial turmoil – evidence from the global financial crisis. *World Dev.* **51**(11), 71–90 (2013)
- Zamore, S., Beisland, L.A., Mersland, R.: Excessive focus on risk? Non-performing loans and efficiency of microfinance institutions. *Int. J. Financ. Econ.* (2021)



# Factors Affecting the Financial Leverage of Vietnam Businesses

Thi Anh Tuyet Le<sup>1(✉)</sup>, Nhan Truong Thanh Dang<sup>1(✉)</sup>, Van Dan Nguyen<sup>2(✉)</sup>,  
and Van Tung Nguyen<sup>1(✉)</sup>

<sup>1</sup> Banking University of Ho Chi Minh City, Ho Chi Minh City, Vietnam  
{tuyetlta, nhanhtt, tungnv}@buh.edu.vn

<sup>2</sup> University of Labour and Social Affairs (Campus II), Hanoi, Vietnam  
dannv@ldxh.edu.vn

**Abstract.** The study assesses the impact of factors affecting the financial leverage of Vietnamese firms. This paper uses Bayesian linear regression model to estimate the relationship of the data series to examine the impact of the characteristics of Vietnamese firms on financial leverage. The results demonstrated that there are a number of factors that positively affect the financial leverage of firms, including: workforce, annual turnover, and export status, gender of business owners and qualifications of the business owner. However, the age of the business has a negative impact on financial leverage. On the basis of these findings, the paper also suggests a number of policies to use financial leverage more effectively on firm value: (1) large-scale, low-debt firms have may consider increasing the debt ratio to take advantage of the tax shield. (2) financial managers of the enterprise need to establish, forecast and evaluate in detail the company's growth ability in different periods (3) train to improve professional knowledge courses for business owners and staff of economic and financial managers; (4) increasing the gender diversity of board members as a tool to control the level of financial leverage of the company.

**Keywords:** financial leverage · vietnamese firms · enterprise characteristics · influencing factors

## 1 Introduction

For a business, financial leverage is really considered one of the important tools. Whether financial leverage is used more or less, a business also needs this type of leverage. However, financial leverage also has two sides, if the business does not know how to take advantage of it effectively, it will certainly not be able to get the desired profit but also have to bear a large debt. When analyzing financial leverage, studies often use factors that show business characteristics such as business size, liquidity, profitability, corporate income tax, growth opportunities, business risk and company age... (Delcoure (2007); Frank and Goyal (2009); Kayo and Kimura (2011); Joeveer (2013); Lemma and Negash (2013); Muthama et al. (2013); Fathi et al. (2014); Memon et al. (2015); Khemiri and

Noubbigh (2018). However, empirical studies in Vietnam (Pham Tien Minh and Nguyen Tien Dung, 2015a, b; Phan Thanh Hiep, 2016; Vo Minh Long, 2017) apparently seem to focus only on characteristics of firms when analyzing financial leverage of firms in Vietnam. Therefore, in this article, the author also selects a number of factors representing the characteristics of Vietnamese firms to study the current situation of financial leverage of Vietnamese firms.

## 2 Literature Review

### 2.1 The Capital Structure Theory of Modigliani and Miller (M&M Theory)

The theory of the relationship between capital structure and firm value was created by two researchers Franco Modigliani and Merton Miller in 1958 and then further developed in 1963. M&M theory content is expressed in two important propositions: The first is about the value of the firm, and the second is about the cost of capital. These clauses are considered in two cases, with and without corporate tax. In a tax-free environment (Modigliani and Miller, 1958), the value of both levered and unlevered firms is the same. In a taxed environment (Modigliani and Miller, 1963), the value of a leveraged firm is higher than that of an unlevered firm due to the benefit of the tax shield. The M&M theory of capital structure is considered a modern theory that explains the relationship between firm value, cost of capital and the level of debt use of the enterprise. M&M theory has clarified the influence of capital structure on cost of capital and enterprise value.

### 2.2 The Theory of Trade-Off in Static Capital Structure

Based on M&M theory, the trade-off theory is a theoretical development that considers the impact of taxes and the cost of financial distress when explaining the capital structure of firms. The trade-off theory was initiated by Kraus and Litzenberger (1973) and developed by Myers (1977). Accordingly, the optimal capital structure reflects the trade-off between the tax benefits of debt and the cost of financial distress. When a business increases its debt ratio, the cost of financial distress increases due to the increased probability of bankruptcy. At some point, the increased value of the interest tax shield will be offset by the expected bankruptcy costs, or at that point, the costs of financial distress will outweigh the benefits of the tax shield from interest. At this point, the value of the firm begins to decline, and the firm's average cost of capital begins to increase as the firm borrows more debt. The benefit of the tax shield is not enough to offset the cost of financial distress.

The trade-off theory also implies that the benefit from debt is only meaningful to the firm in the case of a tax liability. As a result, businesses with accumulated losses will have very little benefit from the tax shield. In addition, businesses that benefit from tax shields from other sources such as depreciation of fixed assets may receive less benefit from financial leverage. Furthermore, in the case of firms with different corporate income tax rates, the enterprise with the higher tax rate will have a greater incentive to borrow debt. This theory also implies that firms with a higher probability of financial distress will

use less debt than firms with less risk of bankruptcy. Therefore, in a case where other factors are the same, firms with high fluctuations in profit before tax and interest usually borrow at low rates. Thus, the trade-off theory is a vital add-on to the act of completing the modern capital structure theory system when considering capital structure in terms of both costs and benefits instead of only calculating benefits and assuming that costs do not exist as in M&M theory.

### 2.3 Pecking Order Theory

The pecking order theory was studied by Myers and Majluf (1984), who divided funding into internal capital (retained earnings) and external capital (borrowed capital and issued new shares) and explain the order of priority among these sources of capital when firms raise capital. According to Myers and Majluf (1984), firms prefer to use retained earnings over borrowed capital and consider issuing new shares to raise capital as a last resort. It means that, internal capital will be prioritized before considering raising capital from outside. The pecking order theory explains why firms tend to prioritize using internal funds, and if they need to raise more external capital, they will prefer to use borrowed capital first. Issuing new equity is often the last resort when a business has exhausted its debt capacity, which means when there is a threat of financial distress of the business to existing creditors as well as to managers.

## 3 Relevant Empirical Studies

In the world, there are many studies on the impact of corporate characteristics on the financial leverage of firms. For example, Chen (2004) analyzed the impact of firm characteristics on the level of debt use of 88 listed companies in China from 1995 to 2000. Companies with more growth opportunities and holding more tangible assets use higher financial leverage. On the other hand, large-scale firms with high profitability will decrease financial leverage.

Research by Gaud et al. (2005) on listed companies in Switzerland found that firms with larger scale and more tangible assets tend to face a considerable high degree of business risks. At that time, the firms often have higher financial leverage than others. On the other hand, companies with more growth opportunities and more profitability are likely to have lower financial leverage than other companies.

Huang (2006) study on listed companies in China and found that companies with large scale, higher level of institutional ownership, high business risk and more growth opportunities will increase financial leverage. On the other hand, firms with high profitability, high non-debt tax shield, and more tangible assets have lower financial leverage than other firms.

Research by Handoo and Sharma (2014) on listed companies in India shows that businesses with more growth opportunities and holding more tangible assets will increase financial leverage. On the other hand, large-scale companies with high profitability, high debt cost, and high corporate income tax are likely to reduce financial leverage.

Researches on factors affecting financial leverage of firms in recent years have also added more factors leading to the difference between groups of firms such as: characteristics of firms, characteristics of ownership of the enterprise, the duration of its

operation. Therefore, in this study, the authors select the following firm-specific factors to analyze the impact on the firm's financial leverage, include: labor force representing the size of the enterprise; annual turnover of the enterprise; export activities of firms; the number of years of establishment of the business, the gender of the business owner, the qualifications of the business owner.

## 4 Research Methodology

Bayes analysis is a powerful analytical tool for statistical modeling, interpretation of results, and data prediction. Estimation accuracy in Bayesian analysis is not limited by sample size and is not affected by limitations such as autocorrelation, endogeneity, variance of variance encountered by the frequency method. Because the data collected is limited from 2007 to 2015, the authors think that using the Bayesian regression method is appropriate.

Based on the research model of Cortez & Susanto (2012), Hernandez-Nicolas et al. (2015) research data were analyzed using Bayesian linear regression model with specific research model as follows:

$$\ln FL_{it} = \beta_0 + \beta_1 \ln TL_{it} + \beta_2 \ln TR_{it} + \beta_3 ex_{it} + \beta_4 y\_est_{it} + \beta_5 gender_{it} + \beta_6 z1_{it} + \beta_7 z2_{it} + \beta_8 z3_{it} + \epsilon_{it}$$

In which:

$\ln FL$ : Financial leverage from 2007 to 2015., which is measured by the natural logarithm of the ratio between firm liability and equity (Finance Leverage = Liability/Equity).

$\ln TL$ : labor force (full time) from 2007 to 2015 (firm size, representing the firm's size), measured by natural logarithm of firm total labor, is a proxy of firm size.

$\ln TR$ : annual revenue from 2007 to 2015, measured natural logarithm of firm annual revenue.

$ex$ :  $ex$  is a binary variable, representing the export factor, which indicates if firms export or not.  $ex$  equals 1 if firms do exports, otherwise 0.

$y\_est$ : number of years of establishment (firm age, representing the firm's age).

$gender$ : gender is a binary variable representing the firm manager gender; gender equals 1 if the manager is male, otherwise 0.

$z1, z2, z3$  are binary variables, representing the educational level of the manager:

$z1$  equals 1 if the manager got undergraduate or graduate degree, otherwise 0.

$z2$  gets 1 if the manager got vocational training degree, otherwise 0.

$z3$  equals 1 if the manager got training without degree, otherwise 0.

$z1 = z2 = z3 = 0$ , no profession.

$\beta$ : Regression coefficient of the model.

$\epsilon$ : Residual.

$\tau$ : Accuracy of error.



## 5 Results and Discussion

In order to select the appropriate a priori information for a large sample size, the article will analyze the sensitivity through five simulations of normally distributed a priori information from 1 to 10,000 (Kelley, 2010). After that based on the maximum mean for Log BF, Log (ML), P(M/y), and the smallest for DIC, the Bayesian coefficients and Bayesian estimation models will be selected. For post-estimation test for the validity of Bayesian inference, the paper will use convergence diagnostics via such tests as autocorrelation, normal distribution, stationary, and Max Gelman-Rubin  $R_c$  test. Last but not least, to check the robustness of parameter spaces of posterior simulations, the research specify prior means from  $-0.5$  to  $0.5$ .

**Table 1.** Bayesian factor test and model test

	Chan	lnFL			
		Avg DIC	Avg log (ML)	Avg log (BF)	P(M/y)
Simulation 1	3	2.03e + 04	-1.02e + 04	<b>1</b>	<b>0.5295</b>
Simulation 2	3	2.03e + 04	-1.02e + 04	-0.1181	0.4705
Simulation 3	3	2.03e + 04	-1.02e + 04	-9.3583	0.0000
Simulation 4	3	2.03e + 04	-1.02e + 04	-19.8288	0.0000
Simulation 5	3	2.03e + 04	-1.02e + 04	-30.2153	0.0000

Source: Author's calculation

Based on the results of Table 1, model 1 is selected and the estimation results are in Table 2 below.

The regression results show that all 8 research variables representing business characteristics have an impact on the financial leverage of the enterprise. In which, there are 7 variables that positively affect the financial leverage of the business, including: lnTL - Representing the size of the business, lnTR - Annual revenue, ex - Export status, gender - Owner gender business, and z1,z2,z3 - Business owner's qualifications; variable y\_est - The age of the business has a negative impact on financial leverage.

The regression coefficient of variable the size of the business shows that an increase in enterprise size will tend to increase the debt ratio and vice versa. The reason is that large-scale firms often have a reputation of reliability in the market, with more transparent information. Along with that, large-scale firms often have better business capacity, their business activities will be more diversified and they are able to adapt and withstand the economy shocks.. In addition, this group of businesses also has a higher debt repayment capacity and is more reputable with creditors than small-sized firms. When diversifying, larger firms have less abnormal returns and less adverse information. As a result, they will have more opportunities to access more debt at a lower cost. The research results are consistent with the results of some studies such as Dang Thi Quynh Anh (2014), Bui Van Thuy 2020; At the same time, the research results also support the trade-off theory.

The regression coefficient of variable annual revenue shows that an increase in annual revenue will increase the debt ratio of the business and vice versa. Increased revenue

**Table 2.** The estimation results

LnFL						
	Mean	Std. Dev.	MCSE	Median	Equal-tailed	
					[95% Cred. Interval]	
lnTL	.03386	.0310895	.000179	.034046	-.0273731	.0952198
lnTR	.1477241	.0209689	.000122	.1476265	.1067393	.1888392
y_est	-.0165038	.0024844	.000014	-.0165048	-.0214116	-.0116449
gender	-.0650891	.048577	.00028	.0647835	-.0293689	.1604247
ex	.2998549	.0905001	.000523	.3003697	.1224471	.4760841
z1	.1921446	.0748367	.000434	.1920253	.0453907	.3383357
z2	.1280991	.0681699	.000394	.1279234	-.0050448	.2619098
z3	.0032143	.0661918	.000385	.0030917	-.1260709	.1343636
_cons	-4.551073	.2621193	.001522	-4.551397	-5.067472	-4.037669
var	2.77366	.0543453	.000314	2.773039	2.668686	2.881467
Number of obs = 5,268						
Avg acceptance rate = 1						
Avg efficiency: min = .9848						
Max Gelman-Rubin Rc = 1						

*Source: Processing results from software*

represents the business growth. The above results are consistent with the research results of Pham Tien Minh and Nguyen Tien Dung (2015a, b), Bui Van Thuy (2020). The above results can be explained that if businesses have good growth rate, they will earn trust from investors as well as creditors. In addition, businesses with high growth potential in revenue need to raise additional capital to increase investment in assets to match the rapid increase in revenue. However, the endogenous resources of the enterprise from retained earnings do not meet the capital needs, then, according to pecking order theory, the company will increase the use of debt instead of issuing shares to cover capital needs. Therefore, when the revenue of the business increases, the debt ratio will increase.

The regression coefficient of variable export status shows that the status of participating in export activities is positively related to the financial leverage of firms. That is, when firms participating in export activities increase, the debt ratio of firms will increase and vice versa. Entering a foreign market is a decision that show the vision of a growing company. Firms profitability increase and volatility may decrease due to international diversification because markets are often imperfectly correlated (Shapiro, 2013). Moreover, exporting companies often receive in advance or part of their sales due to distances and differences in legal systems between countries (Lisboa, 2017). As a result, these companies typically have more free cash flow. After a certain period of export, firms have easy access to long-term debts because they show sustainability in their operations. This result is also consistent with the results found in the study of Pacheco (2016), Lisbon

(2019) on the positive relationship between exports and financial leverage. Again, the trade-off theory suggests that firms that export more will be less liquidity constrained and, therefore, have less difficulty accessing debt and using it to benefit from savings tax for the firms.

The regression coefficient of variable the age of the business shows that the age of the enterprise has a negative effect on the financial leverage of the firms; i.e. the number of years in operation increases, the debt ratio of the business will decrease and vice versa. The above results show that the longer the operating time, the greater opportunity to earn profits. Therefore, to offset their capital requirements, instead of accessing more new loans, these businesses prioritize using internal accumulated capital, which is the retained earnings, to save cost of capital. The above results are consistent with the study of Tran Viet Dung et al. (2019), and the research results also support the pecking order theory.

The regression coefficient of variable gender shows that the firm manager gender is male will use a higher degree of financial leverage than compared to a female company owner. This result is consistent with the study of Shoham & Bazel (2017). Women react differently than men to a variety of business and financial situations, such as risk taking, competition and conflict; Women exhibit strong and significant differences in risk preference, and women's risk aversion increases in opaque environments such as financial markets (Women have 4 different innate behavioral approaches to business, for example, higher risk aversion) (Croson & Gneezy, 2009). Financial leverage is one of the important factors in determining business risks. Therefore, the fact that a company with a higher percentage of male board members will tend to have a higher debt ratio.

The regression coefficient of variable the educational level of the manager shows that a business owner's qualifications have a positive effect on a company's financial leverage. That is, companies with more qualified owners tend to use more debt and vice versa. The above results are consistent with the results found in the study of Rakhmayil (2015). Greater leverage increases the likelihood of financial distress. Therefore, more skills are needed to successfully manage a company with debt in its capital structure. A knowledgeable, qualified and well-connected employer (executive) can analyze the company's mission, stakeholder positions, production or service progress among the factors, and decide on the optimal capital structure for the company and choose the best financial instruments.

## 6 Conclusion

The use of financial leverage (determining capital structure) has an important effect on the value of a business. Research results have shown that factors belonging to the characteristics of firms have an impact on the level of debt use of firms. Some implications from the research results are as follows: *Firstly*, businesses have the common characteristic of benefiting from the size of the business, that is, the larger the size of the business, the more debt it tends to increase which is easier than small businesses. Therefore, large-scale firms with low debt may consider increasing the debt ratio to take advantage of the tax shield and serve as a premise for business growth and development. *Second*, firms with high growth potential, large revenue scale as well as firms with

export status tend to increase the use of financial leverage. Therefore, planners as well as financial managers of firms need to establish and forecast and evaluate the company's growth ability in detail corresponding to different periods in order to have a good base. It helps to proactively set up a plan to mobilize and use funding sources, including loans, to save costs and improve the efficiency of capital use from debt financing. *Third*, it is necessary to invest in training to improve professional knowledge for business owners as well as the team of financial and economic managers in order to be able to apply modern financial management models, risk measurement and identification models as well as the ability to plan and analyze projects for effective financial decisions. *Finally*, the research results also show that the gender of business owners and managers affects the level of debt use. Therefore, businesses may consider increasing the gender diversity of board members as a tool to control the level of financial leverage of the company.

## References

- Anh, Đ.T.Q., Yen, Q.T.H.: Factors affecting the capital structure of firms listed on the Ho Chi Minh City stock exchange. Ho Chi Minh City (HOSE). *J. Dev. Integr.* **18**(28), 09–10 (2014)
- Chen, J.J.: Determinants of capital structure of Chinese-listed companies. *J. Bus. Res.* **57**(12), 1341–1351 (2004)
- Crosan, R., Gneezy, U.: Gender differences in preferences. *J. Econ. Lit.* **47**(2), 448–474 (2009)
- Cortez, M.A., Susanto, S.: The determinants of corporate capital structure: evidence from Japanese manufacturing companies. *J. Int. Bus. Res.* **11**(3), 121 (2012)
- Delcours, N.: The determinants of capital structure in transitional economies. *Int. Rev. Econ. Financ.* **16**(3), 400–415 (2007)
- Dung, T.V., Thanh, B.D.: Factors affecting the capital structure of firms listed on the Vietnam stock exchange. *J. Bank. Sci. Train.* (226- March 2021) (2019)
- Frank, M.Z., Goyal, V.K.: Capital structure decisions: which factors are reliably important? *Financ. Manag.* **38**(1), 1–37 (2009)
- Fathi, S., Ghandehari, F., Shirangi, S.Y.G.: Comparative study of capital structure determinants in selected stock exchanges of developing countries and Tehran Stock Exchange. *Int. J. Acad. Res. Account. Financ. Manag. Sci.* **4**(1), 67–75 (2014)
- Gaud, P., Jani, E., Hoesli, M., Bender, A.: The capital structure of Swiss companies: an empirical analysis using dynamic panel data. *Eur. Financ. Manag.* **11**(1), 51–69 (2005)
- Handoo, A., Sharma, K.: A study on determinants of capital structure in India. *IIMB Manag. Rev.* **26**(3), 170–182 (2014)
- Hernandez-Nicolas, C.M., Martín-Ugedo, J.F., Mínguez-Vera, A.: The influence of gender on financial decisions: evidence from small start-up firms in Spain. *Bus. Adm. Manag.* **18**(4), 93–107 (2015). <https://doi.org/10.15240/tul/001/2015-4-007>
- Huang, G.: The determinants of capital structure: evidence from China. *China Econ. Rev.* **17**(1), 14–36 (2006)
- Jõeveer, K.: What do we know about the capital structure of small firms? *Small Bus. Econ.* **41**(2), 479–501 (2013)
- Kayo, E.K., Kimura, H.: Hierarchical determinants of capital structure. *J. Bank. Financ.* **35**(2), 358–371 (2011)
- Kelley, K.: A Priori Monte Carlo Simulation. In: Salkind, N.J. (ed.), *Encyclopedia of Research Design*, pp. 1–2. SAGE Publication Inc., Thousand Oaks (2010)
- Khémiri, W., Noubigh, H.: Determinants of capital structure: evidence from sub-Saharan African firms. *Q. Rev. Econ. Financ.* **70**, 150–159 (2018)

- Kraus, A., Litzenberger, R.H.: A state-preference model of optimal financial leverage. *J. Financ.* **28**(4), 911–922 (1973)
- Lemma, T.T., Negash, M.: Institutional, macroeconomic and firm-specific determinants of capital structure: the African evidence. *Manag. Res. Rev.* (2013)
- Long, V.M.: The relationship between capital structure and firm value: the case of a company listed on the Ho Chi Minh City Stock Exchange (HSX). *Sci. J. Ho Chi Minh Open Univ.* **12**(1), 180–192 (2017)
- Lisboa, I.: Capital structure of exporter SMEs during the financial crisis: evidence from Portugal. *Eur. J. Manag. Stud.* **22**(1), 25–49 (2017)
- Lisboa, I.: Capital structure choices and exports: the case of the Portuguese mold industry. *Australas. Account. Bus. Financ. J.* **13**(4), 23–45 (2019)
- Minh, P.T., Dung, N.T.: Factors influencing capital structure of Vietnam's real estate enterprises: a move from static to dynamic models. *J. Econ. Dev. (JED)* **22**(4), 76–91 (2015a)
- Modigliani, F., Miller, M.H.: The cost of capital, corporation finance and the theory of investment. *Am. Econ. Rev.* **48**(3), 261–297 (1958)
- Modigliani, F., Miller, M.H.: Corporate income taxes and the cost of capital: a correction. *Am. Econ. Rev.* **53**(3), 433–443 (1963)
- Muthama, C., Mbaluka, P., Kalunda, E.: An empirical analysis of macro-economic influences on corporate capital structure of listed companies in Kenya. *J. Financ. Invest. Anal.* **2**(2), 41–62 (2013)
- Memon, P.A., Rus, R.B.M., Ghazali, Z.B.: Firm and macroeconomic determinants of debt: Pakistan evidence. *Procedia Soc. Behav. Sci.* **172**, 200–207 (2015)
- Minh, P.T., Dung, N.T.: Factors influencing capital structure of Vietnam's real estate firms: a move from static to dynamic models. *J. Econ. Dev. (JED)* **22**(4), 76–91 (2015b)
- Myers, S.C.: Determinants of corporate borrowing. *J. Financ. Econ.* **5**(2), 147–175 (1977)
- Myers, S.C., Majluf, N.S.: Corporate financing and investment decisions when firms have information that investors do not have. *J. Financ. Econ.* **13**(2), 187–221 (1984)
- Pacheco, L.: Capital structure and internationalization: the case of Portuguese industrial SMEs. *Res. Int. Bus. Financ.* **38**, 531–545 (2016)
- Phan, H.T.: Factors affecting the capital structure of industrial firms seen from the GMM model. *J. Financ.* (2016). <https://tapchitaichinh.vn/tai-chinh-kinh-doanh/tai-chinh-doanh-nghiep/nhan-to-anh-huong-den-cau-truc-von-cua-doanh-nghiep-cong-nghiep-nhin-tu-mo-hinh-gmm-109290.html>. Accessed 10 May 2017
- Rakhmayil, S., Yuce, A.: Do CEO qualifications affect capital structure? *J. Appl. Bus. Econ.* **9**(2), 76 (2009)
- Shapiro, A.F.: The business cycle consequences of informal labor markets (Doctoral dissertation, University of Maryland, College Park) (2013)
- Shoham-Bazel, O.: Gender Effects on Firm Capital Structure (Doctoral dissertation, Temple University) (2017)
- Thuy, B.V.: Factors affecting the capital structure of firms listed on the Ho Chi Minh stock exchange. *J. Econ. Forecast.* 126–129 (2020)



# Understanding the Nexus Between Emerging Stock Market Volatility and Gold Price Shocks

Woraphon Yamaka<sup>(✉)</sup>

Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University,  
Chiang Mai, Thailand  
woraphon.econ@gmail.com

**Abstract.** This study investigates the contagion and spillover effects of gold price shocks on the volatility of the Asian emerging stock markets. Gold prices' positive and negative shocks are quantified, and the Vector Autoregressive (VAR) and Copula approaches are employed to measure the spillover and contagion effects between gold price shocks and stock volatilities. Several Copula functions are considered, and the best-fit one is used to explain the correlation or the contagion effect, while the Granger causality test and VAR model are used to examine the casual and spillover effects, respectively. The study's findings show that there is some evidence indicating the volatility spillover, causality, and contagion between gold price shocks and stock volatility.

**Keywords:** Copulas · Contagion · Gold price shocks · Spillover effects · VAR

## 1 Introduction

Emerging stock markets have become increasingly integrated with the global economy in the last few decades. They are in the developing economies that have rapid economic growth, and income per capita per year greater than 15,000 US dollars. The countries where emerging stock markets are operating contain 80% of the world's population, and the size of their combined economy covers 20% of the world economy. It is well known that the stock market is important in the development of emerging economies as the businesses could issue their shares to get money from public investors. Thus, stock markets have been considered a reflection of the economic performance of emerging countries. Many investors have considered this market an alternative, which could provide an excellent opportunity for a higher profit. Although emerging stock markets can offer higher gains to investors due to rapid domestic economic growth, they also expose investors to a greater investment risk due to price volatility, economic and financial uncertainty, and an ongoing global economic slowdown [33]. This indicates that emerging stock markets are as well exposed to various shocks. To solve and mitigate the risk in the stock market, especially during an extreme market downturn, Nguyen et al. [24] and Pastpipatkul, Yamaka, and Sriboonchitta [28] suggested investing in the gold market. Gold is considered a reliable investment and a safe haven commodity that mitigates macroeconomic risk. Therefore, investigating the interconnection between gold price

shocks and the stock market is crucial because this relationship is a relevant reference source for portfolio management and hedging strategies. The relationship between gold prices and stock returns is of significant interest to many scholars in both literature and empirical fields.

Theoretically, there is an inverse relationship between stock market returns and gold prices. There have been circumstances where stock market returns rise and gold prices fall. Gold prices may also rise in sympathy with the fall in stock prices. The reason lies in the perception of investors in the market. Investors who expect a bearish market generally take positions for their investment in gold futures. To reduce the stock market risk, many investors consider gold as the hedging in their portfolios. Gold is a valuable and extremely liquid metal and is classified as a product and a financial asset. Gold has also played an important role as a precious metal with significant portfolio diversification properties [10]. Investors prefer to reconstruct their investment portfolios by replacing some of their stock holdings with gold to protect the losses. Even though there are many empirical papers on stock exchange volatility have been conducted around the globe, few studies have been done on gold return volatility, i.e., the response of gold returns and volatility to public information arrival [18] and the influences of macroeconomic variables on gold returns and volatility [32].

A few studies have shown the impacts of gold price shocks and stock market volatility. However, most of the previous studies have focused on either the co-volatility or stock price and gold price relation. Therefore, this study attempts to provide a new perspective on the stock-gold nexus by identifying gold price shocks by decomposing gold return into positive and negative changes. Then, the relationship between the emerging stock markets and gold price shocks is investigated in various aspects, namely causality, spillover effect, and the risk contagion effect, using the Granger causality test, vector autoregression (VAR) model, and Copula model, respectively. In finance, the terms spillover, contagion, and causality are commonly used interchangeably. The definitions of these three words and their difference are explained in some writings ([14]; Xu [34]; Maneejuk [19]).

The following three aspects mainly reflect the overall contribution of this paper. First, this study applies the Granger causality test and VAR model to examine the causality and spillover effect between gold price shocks and emerging stock markets. It is unclear if emerging stock volatility is being anticipated by gold price shocks or if gold price shocks are just a consequence of the emerging stock volatility. To disentangle these effects, a Granger causality analysis of the stock volatility and different lags and leads of gold price shocks would be very informative. Second, I evaluate the spillover effect between the gold price shocks and emerging stock markets. This spillover can happen in both good and bad times and is not only related to a crisis period. Third, the Copula model is used to describe the nonlinear and asymmetric dependency structure as well as the lower tail coefficients, which can effectively depict the risk of contagion. Using tail coefficients to measure risk contagion from gold price shocks in different emerging markets is also an innovative approach compared with previous studies. It will help the government to provide economic policies from a more macro perspective and investors to make diversified investments with less risk.

This study is organized as follows. Section 2 provides a review of the related study (literature review). Section 3 presents data and methodology. Section 4 presents empirical results and discussion. Section 5 concludes the study.

## 2 Literature Review

In recent years, emerging stock markets have shown substantial growth due to the high capital inflows [4]. However, the emerging stock markets are exposed to global news and events that lead to a risky and uncertain events. The investments in gold are regarded as an inflation hedge, store of value, a source of wealth, and a safe haven asset for stock markets during periods of stock market troubles [2,3].

Following the rapid financialization, the Granger causality, spillover effect, and contagion effect between gold and stock markets have been empirically tested. As for the causality and spillover effect, the most famous and common method to test the causality between two variables is the Granger causality test proposed by Granger [13], and the Vector Autoregressive (VAR) model can be used to detect the linear causality and causal effect between variables. Mishra, Das, and Mishra [22] attempted to investigate the causality and causal effect between gold prices and stock market returns in India and provided evidence of feedback causality between them during 1991–2009. Notably, the gold prices Granger-caused stock market returns, and stock market returns also Granger-caused the gold prices in India during the sample period. Similar to the work of Bhunia and Das [6], they provided support for feedback causality between the selected variables. Their results indicated that the co-movement of gold prices and stock prices is high even during the global financial crisis and after that. However, Hussin et al. [23] that studied the relationship between gold price and the Islamic stock market in Malaysia revealed that gold price is not a valid variable for predicting changes in Islamic stock prices. Choudhry, Hassan, and Shabi [9] investigated co-movements between gold returns and stock market volatility during the global financial crisis in 2007–2008 for the UK, the USA, and Japan. They found that gold may not perform well as a safe haven during the financial crisis period due to the weak bidirectional interdependence between gold returns and stock market volatility. However, gold may be used as a hedge against stock market returns and volatility in the stable financial conditions.

In the VAR framework, Raza, Jawad, Tiwari, and Shahbaz [30] investigated the asymmetric effect of gold prices, oil prices, and their related volatility on emerging stock markets, using monthly data from January 2008 to June 2015. The results showed that gold prices have a negative effect on the stock markets of Mexico, Malaysia, Thailand, Chile, and Indonesia. Additionally, Mensi et al. [21] studied the correlations and volatility spillovers between the S&P 500 and commodity price indices for energy, food, and gold by using a VAR-GARCH model over the period 2000 to 2011, and the results showed a significant transmission in many S&P500 and commodity pairs and found that the highest conditional correlations are the S&P500-gold and the S&P 500-WTI pairs. Hood and Malik [15] studied the role of gold volatility as hedge and safe haven of the US stock market and found that gold is a weak hedge against the US stock market.

As for the measurement of contagion risk, the most classical method is based on the correlation coefficient (Pearson [29]), which only describes a static linear correlation between the variables. Thus, Engle [12] introduced the Dynamic Conditional



Correlation-GARCH model to measure the time-varying correlation between the variables. Later, this model has been receiving increasing interest for researchers and practitioners. Chen and Wang [8] used DCC-GARCH to examine China's dynamic relationships between gold and stock markets. They found that gold acts as a safe haven for only market downturns, while gold does not offer a good risk hedging in market upturns. Basher and Sadorsky [1] considered various DCC-GARCH-type models, namely DCC-GARCH, GO-GARCH, and ADCC-GARCH, to investigate the dynamic correlation. In general, they revealed that gold has a positive correlation with emerging market returns. Thus, gold might not be good hedging for emerging stock markets.

Recently, the DCC-GARCH performance has been questioned by many researchers as there is evidence that the dependencies between financial variables are nonlinear and asymmetric. Therefore, the Copula method has been introduced to measure the contagion or co-movement between two or more variables. Copula was firstly introduced by Sklar [31] and further developed and described by Joe [16]. This model has been widely used to examine the contagion effects between gold and stock. Do, McAleer, and Sriboonchitta [11] studied the impact of gold on the volatility of the emerging ASEAN stock market. Nguyen et al. [24], Pastpipatkul, Yamaka, and Sriboonchitta [26], Pastpipatkul et al. [27] and Beckmann, Berger, and Czudaj [5] used different Copula functions to test the correlation between gold and stock.

Through summarizing the previous literature, there are many methods for testing causality, measuring spillover effects, and quantifying risk contagion, and each method has its own advantages and disadvantages. This study investigates the causality using the Granger causality test. In addition, the VAR model and Copula model are used to explore the spillover effect and contagion effect, respectively. Furthermore, with respect to the impact of gold price on the stock market, all the above studies used either gold price or gold return as a proxy, which may not yield the desired results. In fact, there might be an asymmetric impact from positive and negative gold price shocks on the emerging stock market. Hence, it is of interest to find out whether the output shocks of gold price have different impacts on stock volatility in emerging economies.

### 3 Data and Methodology

#### 3.1 Data

This study uses weekly time series of gold prices and ten emerging stock indexes of Korea (KOR), Thailand (TH), China (CN), Indonesia (ID), India (IND), Vietnam (VN), Philippines (PH), Saudi Arabia (SA), Qatar (QA) and Hong Kong (HK) throughout 1 January 2001, to 31 December 2018. All information every Friday of the week is taken. For some weeks that the market was closed on Friday, the information of Thursday was taken instead. All stock indexes and gold prices are collected from [www.investing.com](http://www.investing.com). Returns of gold and stocks are calculated by taking the first difference of the natural logarithm.

**Table 1.** Descriptive statistics of daily stock market returns and gold prices.

Variable	Mean	Std.Dev	Skewness	Kurtosis	ADF
RGold	0.0033	0.2470	-0.1998	4.5683	-31.1167***
RCN	0.0008	0.0335	-0.0423	5.6511	-27.9771***
RIND	0.0028	0.0298	-0.2767	6.0163	-18.6378***
RID	0.0033	0.0296	-0.6408	7.6383	-31.6757***
RKR	0.0019	0.0301	-0.3639	8.2530	-32.9096***
RHK	0.0010	0.0294	-0.0679	5.5274	-30.4016***
RPH	0.0022	0.0278	-0.1281	7.7969	-31.8466***
RSA	0.0019	0.0339	-0.9987	9.1291	-28.9613***
RQA	0.0028	0.0333	-0.1224	8.3156	-28.4976***
RTH	0.0023	0.0273	-0.9961	10.6266	-12.6105***
RVN	0.0023	0.0398	-0.0702	6.8271	-17.4981***

Notes: ADF is the Augmented Dickey-Fuller unit root test. The null hypothesis is that the variable includes a unit root, and the alternative is that the variable does not include a unit root meaning the variable is stationary. \*\*\* indicates decisive evidence of rejecting the null hypothesis.

Table 1 presents the descriptive statistics of the weekly returns of gold and ten stock markets. It can be seen that the gold price and Indonesian stock return show the highest average return (0.0033), followed by India's stock return and Qatar's stock return, respectively. Gold price has the highest standard derivation (0.2470), followed by Vietnam's and Saudi Arabia's stock returns. This indicates that gold return has enormous fluctuations and is riskier than the emerging stock markets' returns. Negative skewness and high kurtosis values are observed in all market and gold returns, indicating fat tails in the return distributions and non-normality of the data series. The Augmented Dickey-Fuller (ADF) test is employed to examine the stationarity of the data series, and the result indicates that all stock market returns and gold prices are stationary.

### 3.2 Methodology

In order to establish the causality, spillover effect, and risk contagion effect between positive and negative gold price shocks; and emerging stock market volatilities, the study employs a rich set of quantitative techniques such as the Granger causality, VAR model, and various Copula functions. Prior to examining the linkages between the variables, the simple GARCH (1,1) model is used to quantify the conditional variance or volatility of the stock market. Then, the Granger causality test is employed to examine the lead-lag relationship between gold price shocks and stock market volatility. Finally, the VAR and Copula models are used to investigate the spillover effect and contagion effect, respectively.

### 3.2.1 Generalized Autoregressive Conditional Heteroscedasticity (GARCH) Model

In this study, GARCH(1, 1) is used to estimate the volatility of ten emerging stock market returns. Let  $R_{i,t}$  and  $\sigma_{i,t}$  be the return and conditional volatility of stock  $i$  at time  $t$ , respectively. The GARCH(1, 1) for stock  $i$  can be written as

$$R_{i,t} = \mu_i + \varepsilon_{i,t}, \tag{1}$$

where  $\mu_i$  is the constant parameter of the mean equation for stock  $i$ .  $\varepsilon_{i,t}$  is the error term which can be decomposed as follows:

$$\varepsilon_{i,t} = \sqrt{\sigma_{i,t}} z_{i,t}, \tag{2}$$

where  $\sigma_{i,t} = E(\varepsilon_{i,t}^2 | \psi_{i,t-1})$  is the conditional variance of the error and  $z_t \sim \text{skewed-t}(0, 1, df, \gamma)$  is the standardized residual following the skewed-t distribution.  $df$  and  $\gamma$  are degree of freedom and skewness parameters, respectively.  $\psi_{i,t-1}$  is the information set of stock  $i$  available at time  $t - 1$ . According to Bollerslev [7], the conditional variance can be predicted by the lagged conditional variance and the square of the error term in the mean equation. In this study, I consider using GARCH (1, 1) to model the volatility as one lag order can sufficiently capture the volatility clustering of the stock market returns (Oh and Patton [25]). Thus, the conditional variance equation can be written as

$$\sigma_{i,t} = \alpha_{i,0} + \alpha_{i,1} \varepsilon_{i,t-1}^2 + \beta_i \sigma_{i,t-1}, \tag{3}$$

where  $\alpha_{i,0} > 0$ ,  $\alpha_{i,1} > 0$ ,  $\beta_i > 0$  and  $\alpha_i + \beta_i \leq 1$ .

### 3.2.2 Granger Causality Test

According to Granger [13], the Granger causality is a statistical hypothesis test for determining whether one time-series is useful in forecasting another. Note that the causality between positive and negative gold price shocks; and emerging stock market volatilities is examined. Thus, the specific testing equations for this study can be presented as follows:

$$\sigma_{i,t} = \sum_{p=1}^P \phi_p \sigma_{i,t-p} + \sum_{p=1}^P \delta_p g_{t-p} + u_{i,t}, \tag{4}$$

$$g_t = \sum_{p=1}^P \omega_p g_{t-p} + \sum_{p=1}^P \eta_p \sigma_{i,t-p} + v_{i,t}, \tag{5}$$

where  $g_t$  is gold return which is decomposed as the positive and negative shocks ( $g_t^+ = \max(\Delta RGOLD, 0)$  and  $g_t^- = \min(\Delta RGOLD, 0)$ ). To test the causality between stock volatility and gold price shocks in Eqs. 4–5, we set the null hypothesis of no causality as  $H_0 : \delta_1 = \delta_2 = \dots = \delta_p = 0$  ( $H_0 : \eta_1 = \eta_2 = \dots = \eta_p = 0$ ), while the alternative hypothesis is that  $H_0$  is not true. To test this hypothesis, the F-statistic is used.

### 3.2.3 Vector Autoregressive Model (VAR)

The VAR model for each pair  $i$  is formulated as follows:

$$\begin{bmatrix} g_t^+ \\ z_{i,t}^+ \end{bmatrix} = A_{0,i}^+ + \sum_{p=1}^P A_{i,p}^+ \begin{bmatrix} g_{t-p}^+ \\ z_{i,t-p}^+ \end{bmatrix} + e_{i,t}^+, \tag{6}$$

$$\begin{bmatrix} g_t^- \\ z_{i,t}^- \end{bmatrix} = A_{0,i}^- + \sum_{p=1}^P A_{i,p}^- \begin{bmatrix} g_{t-p}^- \\ z_{i,t-p}^- \end{bmatrix} + e_{i,t}^-, \tag{7}$$

where  $A_{i,p}^-$  and  $A_{i,p}^+$  are the autoregressive coefficient or spillover effect between stock  $i$  and  $g_t = (g_t^-, g_t^+)$ .  $A_{0,i}^-$  and  $A_{0,i}^+$  are vectors of the constant term,  $e_{i,t}^+$  and  $e_{i,t}^-$  are error terms which are assumed to follow the normal distribution with mean zero and variance  $\Sigma$ . To select the optimal lag order of a VAR(p) model in Eqs. (6–7), the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are employed, and the best lag is obtained at the lowest AIC and BIC.

### 3.2.4 Copula Approaches

The correlation between gold price shocks and emerging stock market volatilities are measured by the Copula model. Following Sklar’s theorem (Sklar [31]), two continuous marginals can be joined by Copula function  $C(\cdot)$ . Thus, a two-dimensional joint distribution function  $H(x, y)$  can be defined as

$$H(x, y) = C(F_1(x), F_2(y)), \tag{8}$$

where  $F_x(x)$  and  $F_y(y)$  are the cumulative marginal distribution of random variables  $x$  and  $y$ , respectively. If  $F_x(x)$  and  $F_y(y)$  are continuous, the Copula function associated with  $H(\cdot)$  is unique and can be computed by

$$C(u, v) = H(F_1^{-1}(u), F_2^{-1}(v)), \tag{9}$$

where  $F^{-1}(\cdot)$  is the inverse function.  $u$  and  $v$  are uniform  $[0, 1]$  variables, where  $u = F_x(x)$  and  $v = F_y(y)$ . This study aims to find the correlation between the gold price shocks and emerging stock market volatilities; thus, the joint distribution of  $u = F_z(z_{i,t})$  and  $v^+ = F_v(g_t^+/sd(g_t^+))$ , and the joint distribution of  $u = F_z(z_{i,t})$  and  $v^- = F_v(g_t^-/sd(g_t^-))$  are

$$H(u, v^+) = H(F_z(z_{i,t}), F_v(v^+)), \tag{10}$$

$$H(u, v^-) = H(F_z(z_{i,t}), F_v(v^-)), \tag{11}$$

Note that  $sd(\cdot)$  is standard deviation. There are various Copula functions proposed to join the marginal distributions, and the selection of Copula function type is important. In this paper, five Copulas are considered to capture different patterns of dependence between stock market volatility and gold price shocks. Copula functions commonly used in research are Gaussian, Student-t, Gumbel, Clayton, and Frank. Also, AIC is used as the measure for selecting the best-fit Copula function [17]. The Copula specifications are presented in Table 2.

**Table 2.** Copula functions

Copula	Function	Parameter
Gaussian	$C^G(u, v   \theta) = \Phi [\Phi^{-1}(u), \Phi^{-1}(v)]$ .	$\theta = [-1, 1]$
Student-t	$C^S(u, v   \theta) = T_\theta [t_\theta^{-1}(u), t_\theta^{-1}(v)]$ .	$\theta = [-1, 1]$
Gumbel	$C^{Gu}(u, v   \theta_{Gu,t}) = \exp \left( - \left( (-\ln(u))^\theta + (-\ln(v))^\theta \right) \right)^{1/\theta}$	$\theta = [1, \infty]$
Clayton	$C^{Cl}(u, v   \theta) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$ .	$\theta = [0, \infty]$
Frank	$C^{Fr}(u, v   \theta) = -\frac{1}{\theta} \log \left[ 1 + \frac{\exp(-\theta u) - 1}{\exp(-\theta)} \frac{\exp(-\theta v) - 1}{\exp(-\theta)} \right]$ .	$\theta = [-\infty, \infty]$

Notes:  $2 < v \leq 30$  is the degree of freedom of Student-t Copula.  $D$  is (Debye function).

## 4 Empirical Results and Discussion

The series of results are provided in the following sections.

### 4.1 GARCH Model Results

**Table 3.** Estimates from GARCH (1, 1) for the 10 stock markets.

Parameter Estimation	$\mu_i$	$\alpha_{0,i}$	$\alpha_{1,i}$	$\beta_i$	Log-likelihood	Q(10) MBF	ARCH-LM(1)
CN	0.0003 (0.0008)	<0.0001* (<0.0001)	0.1231*** (0.0279)	0.8604*** (0.0284)	1903.939	0.2280	0.8993
IND	0.0033*** (0.0004)	<0.0001 (<0.0001)	0.0987*** (0.0231)	0.8895*** (0.0262)	2022.767	0.4269	0.6320
ID	0.0034*** (0.0008)	<0.0001 (<0.0001)	0.1243*** (0.0326)	0.8481*** (0.0429)	2003.342	0.0792	0.9815
KR	0.0016* (0.0007)	<0.0001* (<0.0001)	0.1286*** (0.0295)	0.8567*** (0.0344)	2044.873	0.5249	0.6271
HK	0.0019* (0.0008)	<0.0001* (<0.0001)	0.0855*** (0.0205)	0.8927*** (0.0248)	2013.785	0.7588	0.8403
PH	0.0028*** (0.0008)	<0.0001 (<0.0001)	0.1081** (0.0344)	0.8568*** (0.0488)	2033.291	0.4216	0.6623
SA	0.0029*** (0.0007)	<0.0001*** (<0.0001)	0.3271*** (0.0490)	0.6348*** (0.0409)	1972.510	0.0060	0.9464
QA	0.0024** (0.0008)	<0.0001** (<0.0001)	0.2589*** (0.0578)	0.6832*** (0.0670)	1960.371	0.1001	0.2883
TH	0.0019** (0.0007)	<0.0001 (<0.0001)	0.0803*** (0.0192)	0.9145*** (0.0199)	2070.848	0.5220	0.3001
VN	0.0002 (0.0008)	<0.0001** (<0.0001)	0.3461*** (0.0634)	0.6741*** (0.0517)	1861.425	0.4001	0.4539

Notes: \*, \*\*, and \*\*\* stand for strong, very strong, and decisive evidence, respectively. ( ) denotes the standard error.

This section presents the estimation results of the GARCH(1, 1) conditional volatility models for ten emerging stock markets. First, the estimation from the GARCH(1, 1) model is presented in Table 3. All the coefficients in the variance equations, i.e., the unconditional volatility ( $\alpha_{0,i}$ ), the ARCH effect ( $\alpha_{1,i}$ ), and the GARCH effect ( $\beta_i$ ) are positive with decisive evidence, indicating that all stock market indices exhibit high volatility persistence. Then, the goodness-of-fit is conducted to test whether the obtained standardized residuals have no autocorrelation and heteroscedasticity. The Ljung-Box Q-statistic at lag 10 and ARCH Lagrange Multiplier (LM) test at lag 1 are used for these proposes. According to the Minimum Bayes factor (MBF) ([20]), there is no autocorrelation and heteroscedasticity of the standardized residuals.

### 4.2 Estimates of Causality Using Granger Causality

**Table 4.** Granger causality test for the relationship between gold price shock and stock market return.

	<i>Stock</i> → <i>PGS</i>		<i>PGS</i> → <i>Stock</i>		<i>Stock</i> → <i>NGS</i>		<i>NGS</i> → <i>Stock</i>	
	MBF-value	Causality	MBF-value	Causality	MBF-value	Causality	MBF-value	Causality
CN	0.3433	No	0.8734	No	0.2093	No	0.3494	No
IND	0.4455	No	0.3394	No	0.0022	Yes	0.6303	No
ID	0.4467	No	0.2094	No	0.0222	Yes	0.2299	No
KR	0.6695	No	0.4033	No	0.0330	Yes	0.4985	No
HK	0.3325	No	0.4983	No	0.0000	Yes	0.3940	No
PH	0.6094	No	0.3098	No	0.2934	No	0.3440	No
SA	0.4950	No	0.5903	No	0.4093	No	0.3904	No
QA	1.0000	No	0.399	No	0.3003	No	0.7445	No
TH	1.0000	No	0.3094	No	0.0203	Yes	0.0000	Yes
VN	0.2343	No	0.5093	No	0.0333	Yes	0.4009	No

Note: *PGS* and *NGS* are, respectively, positive gold price shock and negative gold price shock.

The results of Granger causality tests are presented in Table 4. *Stock* → *PGS* indicates that the stock market volatility Granger causes a positive gold price shock, whereas *PGS* → *Stock* indicates the positive gold price shock Granger causes stock market volatility. Likewise, *Stock* → *NGS* and *NGS* → *Stock* present the Granger causalities between negative gold price shock and stock market volatility. The results of causal relationships between the gold price shocks and the emerging stock market volatilities can be summarized as follows. (1) There is no Granger causality between positive gold price shock and the emerging stock market volatilities. (2) Considering the relationship between negative gold price shock and the emerging stock market volatilities, the unilateral Granger causality is found from stock market volatility to negative gold price for the cases of Indonesia, India, Korea, Hongkong, and Vietnam. (3) The bilateral Granger causality is observed between negative gold price shock and Thai stock market volatility.

4.3 Estimates of Spillover Effect Using the VAR Model

Table 5. VAR model estimation.

POSITIVE GOLD PRICE SHOCKS				NEGATIVE GOLD PRICE SHOCKS			
VARIABLE	$A_{0,i}^+$	$g_{t-1}^+$	$z_{i,t-1}$	VARIABLE	$A_{0,i}^-$	$g_{t-p}^-$	$z_{i,t-1}$
$g_{CN}^+$	0.0101*** (0.0005)	0.0336 (0.0331)	0.0005 (0.0004)	$g_{CN}^-$	-0.0083*** (0.0005)	0.0139 (0.0331)	0.0006 (0.0004)
$z_{CN}^+$	-0.0041 (0.0405)	1.5859 (2.2543)	0.0757* (0.0331)	$z_{CN}^-$	0.0413 (0.0379)	3.3872 (2.2254)	0.0729* (0.0331)
$g_{IND}^+$	0.0101*** (0.0005)	0.0358 (0.0330)	0.0002 (0.0004)	$g_{IND}^-$	-0.0084*** (0.0005)	0.0075 (0.0332)	0.0011* (0.0004)
$z_{IND}^+$	0.0012 (0.0406)	-2.2356 (2.2560)	0.0698* (0.0329)	$z_{IND}^-$	-0.0137 (0.0381)	1.0004 (2.2466)	0.0670* (0.0332)
$g_{ID}^+$	0.0100*** (0.0005)	0.0369 (0.0331)	-0.0001 (0.0004)	$g_{ID}^-$	-0.0084*** (0.0005)	0.0068 (0.0333)	0.0010* (0.0004)
$z_{ID}^+$	-0.0438 (0.0407)	3.5885 (2.2641)	-0.0225 (0.0330)	$z_{ID}^-$	0.0239 (0.0382)	3.5506 (2.2567)	-0.0274 (0.0334)
$g_{KR}^+$	0.0101*** (0.0005)	0.0364 (0.0330)	0.0001 (0.0004)	$g_{KR}^-$	-0.0084*** (0.0005)	0.0102 (0.0331)	0.0011* (0.0004)
$z_{KR}^+$	0.0111 (0.0407)	-1.6421 (2.2595)	-0.0510 (0.0330)	$z_{KR}^-$	-0.0048 (0.0381)	0.1435 (2.2409)	-0.0506 (0.0332)
$g_{HK}^+$	0.0100*** (0.0005)	0.03764 (0.0330)	-0.0005 (0.0004)	$g_{HK}^-$	-0.0084*** (0.0005)	0.0067 (0.0333)	0.0010* (0.0004)
$z_{HK}^+$	-0.0511 (0.0408)	1.7848 (2.2645)	0.0003 (0.0331)	$z_{HK}^-$	-0.0083 (0.0382)	2.8504 (2.2586)	-0.0050 (0.0335)
$g_{PH}^+$	0.0101*** (0.0005)	0.0359 (0.0330)	0.0002 (0.0004)	$g_{PH}^-$	-0.0083*** (0.0005)	0.0141 (0.0334)	0.0003 (0.0004)
$z_{PH}^+$	2.5433 (2.2691)	-0.0167 (0.0330)	-0.0536 (0.0409)	$z_{PH}^-$	-0.0099 (0.0383)	2.0052 (2.2624)	-0.0201 (0.0333)
$g_{SA}^+$	0.0101*** (0.0005)	0.0367 (0.0330)	0.0003 (0.0004)	$g_{SA}^-$	-0.0083*** (0.0005)	0.0171 (0.0330)	0.0006 (0.0004)
$z_{SA}^+$	-0.0352 (0.0404)	1.0363 (2.2450)	0.1050** (0.0329)	$z_{SA}^-$	-0.0158 (0.0379)	1.0041** (2.2158)	0.1045 (0.0329)
$g_{QA}^+$	0.0101*** (0.0005)	0.0360 (0.0330)	-0.0001 (0.0004)	$g_{QA}^-$	-0.0083*** (0.0005)	0.0177 (0.0330)	0.0004 (0.0004)
$z_{QA}^+$	0.0353 (0.0402)	-2.2508 (2.2319)	0.1475*** (0.0327)	$z_{QA}^-$	-0.0031 (0.0376)	-1.7436 (2.2022)	0.1486*** (0.0327)
$g_{TH}^+$	0.0101*** (0.0005)	0.0362 (0.0330)	0.000 (0.0004)	$g_{TH}^-$	-0.0084*** (0.0005)	0.0102 (0.0332)	0.0009* (0.0004)
$z_{TH}^+$	-0.0077 (0.0406)	0.9451 (2.2594)	0.0064 (0.0329)	$z_{TH}^-$	0.0441 (0.0381)	4.9338* (2.2377)	-0.0014 (0.0330)
$g_{VN}^+$	0.0100*** (0.0005)	0.0359 (0.0330)	0.0004 (0.0004)	$g_{VN}^-$	-0.0083*** (0.0005)	0.0173 (0.0330)	0.0008 (0.0004)
$z_{VN}^+$	0.0102 (0.0398)	1.1367 (2.2093)	0.2052*** (0.0324)	$z_{VN}^-$	0.0374 (0.0373)	1.7845 (2.1800)	0.2051*** (0.0324)

Notes: \*, \*\*, and \*\*\* stand for strong, very strong, and decisive evidence, respectively. The parenthesis ( ) denotes the standard error.

The spillover effects between positive/negative gold price shocks and the emerging stock markets are reported in Table 5. The results show strong evidence of positive spillover effects of adverse gold price shocks on Saudi Arabian stocks with a value of 1.0041 and on Thai stocks with a value of 4.9338. For other pairs, there is no evidence

supporting the spillover effect between gold price shocks and stock market volatility. Therefore, this study concludes that the spillover effects between stock market and gold price shocks for emerging stock markets are quite weak; specifically, emerging stock market volatilities do not react to gold shocks, except for the cases of Thai and Saudi Arabian financial markets.

#### 4.4 Estimation Results of Copula

The estimates of dependence between gold price shocks (*PGS* and *NGS*) and ten emerging stock market volatilities are provided in Table 6. Five static Copula functions (Gaussian, Student-t, Clayton, Gumbel, and Frank) are also compared using the AIC. The best model is presented in bold number.

**Table 6.** Estimated Copula parameters and their corresponding AICs

COPULA		CN	IND	ID	KR	HK	PH	SA	QA	TH	VN
GAUSSIAN	$\theta$	0.10 (0.01)	0.01 (0.18)	0.06 (0.19)	-0.04 (0.13)	0.02 (0.20)	0.00 (0.19)	-0.04 (0.04)	-0.02 (-0.01)	0.04 (0.14)	0.04 (-0.01)
	AIC	-2.50 (-2.62)	1.97 (-12.77)	0.48 (-14.24)	1.39 (-5.17)	1.83 (-18.07)	2.00 (-14.45)	1.14 <b>(1.32)</b>	<b>1.70</b> (1.97)	1.07 <b>(-6.53)</b>	1.36 <b>(1.93)</b>
STUDENT-T	$\theta$	0.10 (0.11)	0.03 (0.16)	0.07 (0.16)	-0.01 (0.10)	0.03 (0.19)	0.03 (0.17)	-0.02 (0.02)	-0.02 (0.00)	0.11 (0.14)	0.04 (-0.01)
	AIC	0.42 (-1.04)	<b>-4.19</b> (-12.98)	2.73 (-17.23)	2.16 (-5.35)	0.26 (-18.29)	2.21 (-16.10)	3.41 (2.65)	5.54 (2.72)	v4.85 (-6.10)	7.40 (5.77)
CLAYTON	$\theta$	0.08 (0.08)	0.00 (0.12)	0.01 (0.12)	0.00 (0.09)	0.00 (0.15)	0.01 (0.13)	0.00 (0)	0.00 (0.01)	0.00 (0.07)	0.05 (0)
	AIC	-0.07 (-3.38)	2.00 <b>(-14.38)</b>	1.96 <b>(-17.96)</b>	2.00 <b>(-8.08)</b>	2.00 <b>(-19.44)</b>	1.97 <b>(-18.31)</b>	2.01 (2.00)	2.00 <b>(1.78)</b>	2.00 (-5.24)	1.06 (2.01)
GUMBEL	$\theta$	1.02 (1.05)	1.03 (1.12)	1.02 (1.13)	1.00 (1.10)	1.03 (1.13)	1.00 (1.11)	1.01 (3.35)	1.01 (1.00)	1.05 (1.11)	1.00 (0)
	AIC	0.50 (-0.12)	-0.61 (-8.98)	1.58 (-10.44)	2.01 (-5.53)	<b>-1.44</b> (-11.63)	2.01 (-8.00)	2.01 <b>(0.51)</b>	2.00 (2.00)	-1.20 (-7.18)	2.02 (2.00)
FRANK	$\theta$	0.82 (1.06)	0.32 (0.93)	0.90 (1.21)	0.28 (0.42)	0.35 (1.19)	0.51 (1.23)	0.14 (-0.15)	-0.16 (0.11)	0.94 (0.87)	0.44 (-0.07)
	AIC	<b>-4.97</b> <b>(-6.49)</b>	0.82 (-6.10)	<b>-5.10</b> (-7.30)	<b>1.16</b> (0.64)	0.53 (-12.02)	<b>-0.84</b> (-10.38)	1.84 (1.89)	1.76 (1.93)	<b>-6.93</b> (-3.52)	<b>0.08</b> (1.97)

Notes: The parenthesis () presents the dependence parameter and AIC of negative gold price shock- stock market volatility nexus. The bold number indicates the best-fit Copula function.

According to Table 6, the Frank Copula is the most appropriate function for joining the positive gold price shock and stock markets of China, Indonesia, Korea, Philippines, Thailand, and Vietnam. The Gaussian Copula is an appropriate function for joining positive gold price shocks and Saudi Arabian and Qatar markets. For the dependence between negative gold price shock and emerging stock markets, Clayton is the best-fit Copula function for paring adverse gold price shocks and Indian, Indonesian, Korean, Hong Kong, and Philippines markets, while Gumbel is the best-fit Copula for joining negative gold price shock and Saudi Arabian and Thai markets. The best-fit Copulas are then reported in Table 7.



Overall, the dependence parameters of negative gold price shock and stock-market pairs are mostly higher than the positive price shock and stock-market pairs. Among all pairs of gold price shocks and stock market, the result of Kendall's tau reveals that the negative price shock and Saudi Arabian market pair presents the highest degree of correlation, while the lowest correlation is found in the case of negative price shock and Vietnam's market pair. Considering the tail dependence, the upper tail dependence is found in the cases of Hongkong-PGS, Saudi Arabia-NGS, and Thailand-NGS pairs. It is quite interesting that the negative price shock performs a high correlation with Saudi Arabian and Thai markets during the bullish regime, implying that the large drop in the gold price could decrease the volatilities of Saudi Arabian and Thai markets during the market upturn regime. Therefore, the investors of these two countries should be aware of the negative shock of gold prices. Regarding the lower tail dependence, which reflects the degree of contagion, it is found that there exists a weak degree of contagion between negative gold price shock and Hong Kong stock market volatility. This is another interesting result as both forms of gold price shocks have presented a degree of tail dependence on Hong Kong. This indicates that Hong Kong stock return volatility is quite sensitive to the gold price shocks during extreme events.

**Table 7.** Dependence measure between gold price shocks and stock return

MARKET	SELECTED COPULA	DEPENDENCE PARAMETER	KENDALL'S TAU	UPPER TAIL	LOWER TAIL
CH	Frank (Frank)	0.82 (1.06)	0.09 (0.12)	0 (0)	0 (0)
IND	Student-t (Clayton)	0.03 (0.12)	0.02 (0.06)	0 (0)	0 (0)
ID	Frank (Clayton)	0.90 (0.12)	0.10 (0.06)	0 (0)	0 (0)
KR	Frank (Clayton)	0.28 (0.09)	0.03 (0.04)	0 (0)	0 (0)
HK	Gumbel (Clayton)	1.03 (0.15)	0.03 (0.07)	0.04 (0)	0 (0.01)
PH	Frank (Clayton)	0.51 (0.13)	0.06 (0.06)	0 (0)	0 (0)
SA	Gaussian (Gumbel)	-0.04 (-3.35)	-0.03 (-0.34)	0 (0.22)	0 (0)
QA	Gaussian (Gaussian)	-0.02 (0.04)	-0.02 (0.03)	0 (0)	0 (0)
TH	Frank (Gumbel)	0.94 (1.11)	0.10 (0.10)	0 (0.13)	0 (0)
VN	Frank (Gaussian)	0.44 (-0.01)	0.05 (-0.01)	0 (0)	0 (0)

Notes: The parentheses ( ) denotes the result of negative gold price shock and stock market volatility.

## 5 Conclusion

This study aims to examine the causality, spillover, and correlation effects between gold price shocks and ten Asian emerging stock markets. Three econometric approaches, namely the Granger causality test, VAR, and Copula, are conducted to achieve this research goal. These methods reveal three key findings. First, the Granger causality test reveals weak evidence supporting the causality relation between stock volatility and positive gold price shock. However, some slight evidence shows that six out of ten emerging stock markets can be viewed as the predictor of the negative price shock, while only Thai stock is predicted by negative price shock. Second, the VAR estimation results also revealed a weak spillover between stock and gold price shocks. Third, several Copula functions are compared, and the best-fit Copula is used to reveal the degree of dependence as well as tail dependence. The results show that the correlation between stock return and gold price shock is not so high. Likewise, the degree of tail dependence is observed in some pairs. Finally, results from this investigation reveal some robust similarities in such a way that there is a low degree in the stock-gold price nexus in many perspectives.

For the further study, more linkage dimensions, such as quantile to quantile correlation and dynamic conditional correlation, are suggested in order to confirm the relationship between gold price shocks and stock markets.

## References

1. Basher, S.A., Sadorsky, P.: Hedging emerging market stock prices with oil, gold, VIX, and bonds: a comparison between DCC, ADCC and GO-GARCH. *Energy Econ.* **54**, 235–247 (2016)
2. Baur, D., Lucey, B.: Is gold a hedge or a safe haven? An analysis of stocks. *Bonds and Gold. SSRN Electron. J.* **40** (2009). <https://doi.org/10.2139/ssrn.952289>
3. Baur, D.G., McDermott, T.K.: Is gold a safe haven? International evidence. *J. Bank. Financ.* **34**(8), 1886–1898 (2010)
4. Beckmann, J., Berger, T., Czudaj, R.: Does gold act as a hedge or a safe haven for stocks? A smooth transition approach. *Econ. Model.* **48**, 16–24 (2015)
5. Beckmann, J., Berger, T., Czudaj, R.: Tail dependence between gold and sectorial stocks in China: perspectives for portfolio diversification. *Empir. Econ.* **56**(3), 1117–1144 (2019)
6. Bhunia, A., Das, A.: Association between gold prices and stock market returns: empirical evidence from NSE. *J. Exclusive Manage. Sci.* **1**(2), 1–7 (2012)
7. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **31**(3), 307–327 (1986)
8. Chen, K., Wang, M.: Does gold act as a hedge and a safe haven for China's stock market? *Int. J. Financ. Stud.* **5**(3), 18 (2017)
9. Choudhry, T., Hassan, S.S., Shabi, S.: Relationship between gold and stock markets during the global financial crisis: evidence from nonlinear causality tests. *Int. Rev. Financ. Anal.* **41**, 247–256 (2015)
10. Ciner, C.: On the long run relationship between gold and silver prices A note. *Glob. Financ. J.* **12**, 299–303 (2001). [https://doi.org/10.1016/S1044-0283\(01\)00034-5](https://doi.org/10.1016/S1044-0283(01)00034-5)
11. Do, G., McAleer, M., Sriboonchitta, S.: Effects of international gold market on stock exchange volatility: evidence from ASEAN emerging stock markets. *Econ. Bull.* **29**, 599–610 (2009)

12. Engle, R.: Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **20**(3), 339–350 (2002)
13. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. *Econom.: J. Econom. Soc.* 424–438 (1969)
14. Gulzar, S., Mujtaba Kayani, G., Xiaofeng, H., Ayub, U., Rafique, A.: Financial cointegration and spillover effect of global financial crisis: a study of emerging Asian financial markets. *Econ. Res.-Ekonomiska Istraživanja* **32**(1), 187–218 (2019)
15. Hood, M., Malik, F.: Is gold the best hedge and a safe haven under changing stock market volatility? *Rev. Financ. Econ.* **22**(2), 47–52 (2013)
16. Joe, H.: *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, Boca Raton (1997)
17. Kaewsompong, N., Maneejuk, P., Yamaka, W.: Bayesian estimation of archimedean Copula-based sur quantile models. *Complexity* **2020**, 1–15 (2020)
18. Kutan, A., Aksoy, T.: Public information arrival and gold market returns in emerging markets: evidence from the Istanbul gold exchange. *Sci. J. Adm. Dev.* **2**, 13–26 (2004)
19. Maneejuk, P., Yamaka, W.: Predicting contagion from the US financial crisis to international stock markets using dynamic Copula with google trends. *Mathematics* **7**(11), 1032 (2019)
20. Maneejuk, P., Yamaka, W.: Significance test for linear regression: how to test without P-values? *J. Appl. Stat.* **48**(5), 827–845 (2021)
21. Mensi, W., Beljid, M., Boubaker, A., Managi, S.: Correlations and volatility spillovers across commodity and stock markets: linking energies, food, and gold. *Econ. Model.* **32**, 15–22 (2013)
22. Mishra, P.K., Das, J.R., Mishra, S.K.: Gold price volatility and stock market returns in India. *Am. J. Sci. Res.* **9**(9), 47–55 (2010)
23. Hussin, M.Y.M., Muhammad, F., Razak, A.A., Tha, G.P., Marwan, N.: The link between gold price, oil price and Islamic stock market: experience from Malaysia. *J. Stud. Soc. Sci.* **4**(2), 161–182 (2013)
24. Nguyen, C., Bhatti, M.I., Komorníková, M., Komorník, J.: Gold price and stock markets nexus under mixed-Copulas. *Econ. Model.* **58**, 283–292 (2016)
25. Oh, D.H., Patton, A.J.: Time-varying systemic risk: evidence from a dynamic Copula model of CDS spreads. *J. Bus. Econ. Stat.* **36**(2), 181–195 (2018)
26. Pastpipatkul, P., Yamaka, W., Sriboonchitta, S.: Co-movement and dependency between New York stock exchange, London stock exchange, Tokyo stock exchange, oil price, and gold price. In: Huynh, V.N., Inuiguchi, M., Demoeux, T. (eds.) *IUKM 2015. LNCS*, vol. 9376, pp. 362–373. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-25135-6\\_34](https://doi.org/10.1007/978-3-319-25135-6_34)
27. Pastpipatkul, P., Yamaka, W., Sriboonchitta, S.: Analyzing financial risk and co-movement of gold market, and Indonesian, Philippine, and Thailand stock markets: dynamic Copula with Markov-switching. In: Huynh, V.N., Kreinovich, V., Sriboonchitta, S. (eds.) *Causal Inference in Econometrics. Studies in Computational Intelligence*, vol. 622, pp. 565–586. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-27284-9\\_37](https://doi.org/10.1007/978-3-319-27284-9_37)
28. Pastpipatkul, P., Yamaka, W., Sriboonchitta, S.: Portfolio selection with stock, gold and bond in Thailand under vine Copulas functions. In: Anh, L., Dong, L., Kreinovich, V., Thach, N. (eds.) *ECONVN 2018. Studies in Computational Intelligence*, vol. 760, pp. 698–711. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-73150-6\\_55](https://doi.org/10.1007/978-3-319-73150-6_55)
29. Pearson, K.: VII. Note on regression and inheritance in the case of two parents. *Proc. Roy. Soc. London* **58**(347–352), 240–242 (1895)
30. Raza, N., Shahzad, S.J.H., Tiwari, A.K., Shahbaz, M.: Asymmetric impact of gold, oil prices and their volatilities on stock prices of emerging markets. *Resour. Policy* **49**, 290–301 (2016)
31. Sklar, M.: Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–231 (1959)

32. Tully, E., Lucey, B.M.: A power GARCH examination of the gold market. *Res. Int. Bus. Financ.* **21**(2), 316–325 (2007)
33. Wen, X., Cheng, H.: Which is the safe haven for emerging stock markets, gold or the US dollar? *Emerg. Mark. Rev.* **35**, 69–90 (2018)
34. Xu, G., Gao, W.: Financial risk contagion in stock markets: causality and measurement aspects. *Sustainability* **11**(5), 1402 (2019)



# How Does Energy Consumption Matter for Economic Growth? A Bayesian Data Analysis

Nguyen Ngoc Thach<sup>1</sup> (✉) and Phan Thi Lieu<sup>2</sup>

<sup>1</sup> Banking University HCMC, HCMC, 36 Ton That Dam Street, District 1, Ho Chi Minh City, Vietnam

thachnn@buh.edu.vn

<sup>2</sup> University of Labour and Social Affairs HCMC, HCMC, 1018 To Ky Street, District 12, Ho Chi Minh City, Vietnam

**Abstract.** Energy is traditionally regarded to significantly contribute to most economic activities. Countries select the type of energy most appropriate for their needs based on geographical conditions and capital, technology, and labor availability. However, most earlier studies used various frequentist methods, which produced so far different findings. This study focuses on clarifying the effects of renewable and non-renewable energy on economic growth in the ASEAN-5 countries: Malaysia, Philippines, Singapore, Thailand and Vietnam. By unitizing a Bayesian Markov chain Monte Carlo (MCMC) sampling algorithm on a panel data sample for 1990–2019, the estimation results provide new evidence on the contribution of energy consumption to economic growth in the investigated nations. As a result, non-renewable energy has a positive impact on economic growth while renewable energy exerts the opposite effect. That is because of an extensive, energy-intensive growth model followed by the ASEAN countries, in which non-renewable energy is an important engine of economic growth.

**Keywords:** Bayesian MCMC algorithm · renewable energy · non-renewable energy · economic growth

## 1 Introduction

Economic growth is the aspiration of every nation. Particularly, energy is seen as a crucial component of manufacturing and is crucial for economic expansion in many parts of the world (Saldivia et al., 2020). However, it is also recognized that energy consumption, particularly the use of non-renewable energy sources, is the main driver of environmental degradation issues that have a detrimental impact on human health (Agbede et al., 2021). It is not easy to balance, calculate, and select the appropriate type of energy for economic growth. It is even more difficult for countries with low per capita income, mainly manufacturing outsourcing, limited capital and technology. Therefore, it is essential to comprehend the function that energy consumption plays in economic activity in order to establish policies and make informed decisions (Stern, 2000).

The ASEAN region is predicted to become the fourth largest economy of the world by 2030. One of the most important requirements for achieving this growth goal is continuous access to energy. Due to an extensive, natural resource-intensive growth model, ASEAN's electricity demand is growing at a six percent annual rate. According to the International Energy Association (IEA), overall demand has increased by 80% since 2000, which is expected to increase by 60% between now and 2040. Despite it being one of the fastest growing regions in the world, renewable energy supply only meets 15% of the needs of these regional countries, the rest comes from non-renewable energy sources. This is really a great challenge to the ASEAN nations in terms of a probable energy crisis during economic expansion, as well as issues with environmental degradation.

As a consequence of the aforementioned circumstance, the study focuses on predicting the contribution of renewable energy and non-renewable energy consumption to economic growth of the ASEAN countries with an expectation that non-renewable energy consumption positively affects economic growth, while renewable energy exerts the opposite effect. The main findings of the study support basic theories and provide a valid empirical foundation for growth policymakers.

## 2 Literature Review

### 2.1 Theories of Growth

In analyzing determinants affecting economic growth, the Cobb–Douglas production function is of great interest. However, this model contains only two main variables: capital and labor. Robert Solow is the first researcher who extended the growth model including technology as an exogenous variable. In contrast to Solow, Paul Romer considered technology as an endogenous variable in his growth model. Recently, Amin and Alam (2018) argued that technology is thought to be related to energy, and therefore energy plays an indirect role in production. Developing this idea, Han et al. (2019) proposed the Coordinated Development of Energy Structure and Industrial Structure (E&I-SD) model for China. This E&I-SD model may serve as a guidance to support policymakers in their efforts to coordinate the development of energy structure and industrial structure strategies (Han et al., 2019).

### 2.2 Empirical studies

The connection between energy consumption and economic growth has also attracted an increasing attention from many applied researchers. Empirical results are often different, even conflicting. For example, using the panel data of 26 OECD countries from 1971 to 2014, Tran et al. (2022) reported a non-linear relationship between economic growth and energy consumption. They explored the threshold effect of GDP on the causality between GDP and energy consumption. The threshold regression technique was applied to check whether a GDP threshold exists in the relationship between GDP and energy consumption, and the Granger causality test based on panel VECM is used to test causality across GDP regimes. The empirical findings highlight the existence of a

GDP threshold at which the effects of GDP on energy consumption and the direction of the energy-growth causality relies on the initial GDP value. When real GDP per capita is less than US\$ 48,170, there is unidirectional causality from energy consumption to GDP in both the short- and long run. With GDP larger than US\$ 48,170 per capita, the Granger analysis results indicate that there is no immediate cause-and-effect link between GDP and energy consumption; GDP is found not to impact energy consumption in the long run. Yang et al. (2021) sought how various sources of energy can influence the output level in Pakistan. The results showed that if fossil energy consumption contributes to economic growth, energy consumption negatively affects the output growth of Pakistan. This confirms that fossil fuel has an inverted U-shaped effect on output per person in the long- as well as short-run in Pakistan. We can see that renewable energy usage may reduce economic growth in the early stages, but doubling the level of renewable energy use will make the Pakistan economy grow significantly. As a result, an interesting finding is that fossil fuels can boost economic growth in the early stages, while renewables drive economic growth in the long run. Unlike Yang et al. (2021), Saldivia et al. (2020) provided mixed evidence on the relationship between energy consumption and GDP for different time horizons and subgroups among the U.S. states.

In addition to the general impact of energy consumption on economic growth, the effect of different energy sources has also drawn attention. Dabboussi and Abid (2022) explored the threshold effect of sectoral renewable energy consumption on economic growth investigating the non-linear relationship between economic growth, capital, labor, and sectoral renewable energy consumption in the United States between 1981Q1 and 2021Q1. By using a threshold detection method, they showed that the effect of sectoral renewable energy consumption on economic growth is positive in case electric power, industrial, residential, commercial, and transportation exceed a renewable energy consumption threshold of  $\ln\text{REC} > 7.095$ ,  $\ln\text{REC} > 6.138$ ,  $\ln\text{REC} > 4.897$ ,  $\ln\text{REC} > 3.212$ , and  $\ln\text{REC} > 2.849$  respectively. Indeed, for the renewable energy investments to positively contribute to the United States' economic growth, each sector must exceed a certain level of renewable energy consumption. In case the United States uses renewable energy below a certain threshold, the impact on economic growth is negative. However, on the contrary, Sasana and Ghazali (2017) found that renewable energy negatively affects economic growth in BRICS countries. Awodumi and Adewuyi (2020) examined the role of non-renewable energy in economic growth and carbon emissions in Africa's top oil producing economies from 1980 to 2015, by using non-linear autoregressive distributed lag (NARDL) technique. The results showed that non-renewable energy consumption has both positive and negative impacts in different countries. Long-run results demonstrated that changes in natural gas consumption per capita had a significant negative effect on real GDP per capita in Gabon, Angola and Egypt, but significant positive effect in Nigeria.

In sum, the overviewed previous studies on the same topic focused on the economies beyond the ASEAN region using out-of-date frequentist estimators, which cannot provide probability statements and interpret non-significant p-values in the case of the power failure. Furthermore, empirical outcomes are different or contradictory. Thus, the current research addresses the ASEAN-5 countries to offer new evidence through the use of a Bayesian MCMC sampler.

### 3 Methodology

The article focuses on analyzing the effects of renewable and non-renewable energy consumption on economic growth in ASEAN-5 countries. Based on the research of Yang et al. (2021), the authors propose an econometric model as follows:

$$\text{GDP}_{it} = \beta_0 + \beta_1 \text{REC}_{it} + \beta_2 \text{EC}_{it} + \beta_3 \text{LF}_{it} + \beta_4 \text{EVI}_{it} + u_{it}.$$

where  $u$  is the error terms,  $t$  is the years (from 1990 to 2020);  $i$  are the countries (Malaysia, Philippines, Singapore, Thailand, Vietnam). Income per capita is chosen as a proxy for economic growth. Due to skew distribution in the data, the variables are logarithmic to smooth the data (Table 1).

**Table 1.** Variables, units and data sources

Variable	Variable explanation	Unit	Sources
GDP	GDP per capita	constant 2015 US\$	World Bank
REC	Renewables per capita	kWh	Our World in Data
EC	Fossil fuels per capita	kWh	Our World in Data
LF	Labor force, total	Person	World Bank
EVI	Export value index (2000 = 100)	%	World Bank

Sources: Author

As recommended by many theoretical and empirical studies on the advantages of the Bayesian way over the frequentist framework (Anh et al., 2018; Hung T. Nguyen and Thach, 2018; Hung T. Nguyen et al., 2019; Kreinovich et al., 2019; Thach, 2020), a Bayesian MCMC simulation algorithm, namely the Metropolis–Hastings sampler is employed to analyze the impact of renewable and non-renewable energy on economic growth in the researched countries.

## 4 Empirical Results

### 4.1 Bayesian Simulation Results

In this article, the authors employ a Bayesian regression with informative normal priors of (0,1) for the structural parameters, keeping the prior on the overall variance uninformative. The results are presented in Table 3. However, let us check MCMC convergence before proceeding to inference. Once MCMC chains have converged to the stationary distribution, estimation results will be useful for further analysis. Among various convergence tests, popular Gelman-Grubin diagnostic provides an important indicator of convergence, so we will use this tool. As the Gelman-Grubin test result exhibits, max Rc equals 1.000134, satisfying convergence rule ( $Rc < 1.1$ ) (Table 2). Hence we can conclude that MCMC chains have converged for all the parameters of the model.

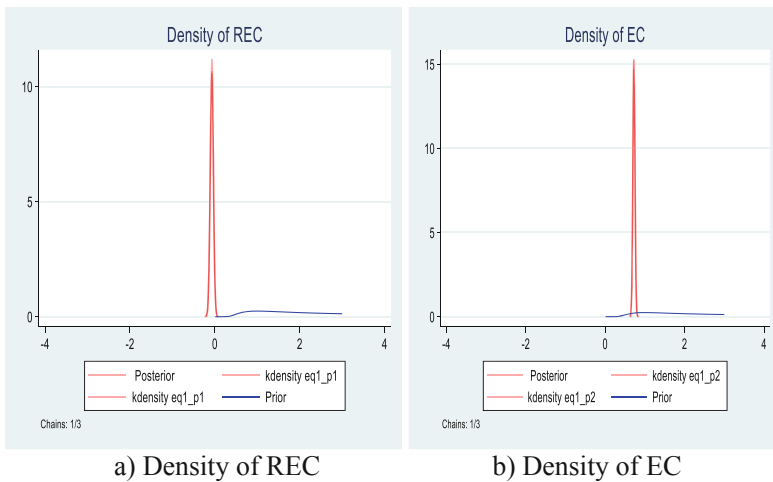


**Table 2.** Gelman-Rubin convergence diagnostic

Variable	Rc
GDP	
REC	1.000014
EC	1.00005
LF	1.000084
EVI	0.9999798
intercept	0.9999568
sigma2	1.000134

Source:the authors’ calculation

In order to access how prior specifications influence the posteriors, we generate prior-posterior plots for the parameters of interest, REC and EC. As seen from Fig. 1, the posterior density of both REC and EC greatly differs from their prior density. The result means that the priors on their parameters have little effect on the posteriors. That is because a large data sample is used in our model.



**Fig. 1.** Prior and posterior density. Source: the authors’ calculation

According to the obtained outcomes, renewable and non-renewable energy consumption exerts contradictory impacts on economic growth in the 5 ASEAN countries with renewable energy consumption negatively influencing and non-renewable energy consumption positively affecting economic growth. This result is similar to the studies of Sasana and Ghozali (2017) or Awodumi and Adewuyi (2020) in the context of Gabon, Angola and Egypt, while differs from those of the other studies reviewed above. The authors believe that this result comes from the ASEAN countries’ over-reliance on

**Table 3.** Bayesian posterior estimates

Variable	Reg. coef	Std. dev	MCSE	95% credibility interval
GDP				
REC	-0.072	0.037	0.000	-0.146/0.000
EC	0.724	0.026	0.000	0.673/0.776
LF	-0.147	0.038	0.000	-0.222/-0.072
EVI	0.542	0.173	0.001	0.205/0.880
intercept	-0.090	0.996	0.005	-2.053/1.847
sigma2	0.059	0.007	0.000	0.047/0.074

Source: the authors' calculation

non-renewable energy sources in their economic development process (85 percent of total energy used). Meanwhile, renewable energy sources have been underutilized due to high costs of exploitation, unfavorable natural conditions, etc. In addition, the export value has a positive effect on per capita income in the considered countries, while labor force is negatively correlated to GDP growth.

## 4.2 Interpreting Bayesian Simulation Results

Because of the harmful effects of fossil fuels on the environment, many countries around the world, including the ASEAN-5 countries, have strengthened policies on the use of renewable energy. This represents the contribution of both renewable and non-renewable energy in boosting GDP growth. However, the results also indicate that, despite the harmful effects from using fossil fuels, the countries in this region still critically depend on them. Projects using renewable energy usually require big investments. This is seen as an enormous barrier for developing countries.

## 5 Conclusion and Policy Implications

### 5.1 Conclusion

The purpose of the study is to explore the role of renewable and non-renewable energy consumption in the economic growth of the ASEAN-5 countries for period 1990–2019. By applying a MCMC sampling algorithm within the Bayesian framework, the empirical results of the study show some important points as follows:

Firstly, renewable energy has a negative impact on economic growth; increasing the share of renewable energy consumption causes a slight decrease in GDP per capita.

Secondly, non-renewable energy has a strong positive impact on economic growth in the ASEAN countries of interest.

Thirdly, while the export value index shows a strong positive link with economic growth, labor force has a negative impact on growth.

## 5.2 Policy Implications

The findings show that renewable energy has a negative impact on economic growth. This is not to say that countries should exclude renewable energy from economic activities. The usage of renewable energy has good effects on the environment, which is a requirement for nations to achieve their sustainable development goals. Fossil fuels, which increase greenhouse gas emissions, are one of the biggest causes of pollution (Bekun et al., 2019; Salari et al., 2021). The issue that the ASEAN nations must solve is how to increase the proportion of renewable energy in economic activity and decrease reliance on fossil fuels. In order to address this issue, the authors suggest the following measures:

Firstly, it is necessary to issue a common legal framework for renewable energy development. According to the experience of countries that succeeded in developing renewable energy such as Germany, China, India and European and American countries, there is a need to perfect regulations related to the use of renewable energy. As a result, the legislative framework for the development of renewable energy will be beneficial.

Secondly, a suitable electricity price must be created for electricity generated by other renewable energy sources as well as by solar and wind energy. To encourage openness and competitive fairness in the power trading market, cut back on subsidies for electricity produced from fossil fuels.

Thirdly, the establishment of a short-, medium-, and long-term strategy for the growth of renewable energy is required, with defined goals for each stage of economic development.

Fourthly, there should be mechanisms and policies in place to encourage investors, particularly foreign investors, such as corporate income tax exemption/reduction and land rent reduction. This will assist the ASEAN countries in attracting capital and capitalizing on advanced countries' high-tech technical levels in the field of renewable energy development.

## References

- Agbede, E.A., Bani, Y., Azman-Saini, W.N.W., Naseem, N.A.M.: The impact of energy consumption on environmental quality: empirical evidence from the MINT countries. *Environmental Science and Pollution Research* **28**(38), 54117–54136 (2021, 2021/10/01). <https://doi.org/10.1007/s11356-021-14407-2>
- Amin, S.B., Alam, T.: The relationship between energy consumption and sectoral output in bangladesh: an empirical analysis. *The Journal of Developing Areas* **52**(3), 39–54 (2018). <https://doi.org/10.1353/jda.2018.0035>
- Anh, L.H., Le, S.D., Kreinovich, V., Thach, N.N. (eds.): *Econometrics for Financial Applications*. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-73150-6>
- Awodumi, O.B., Adewuyi, A.O.: The role of non-renewable energy consumption in economic growth and carbon emission: Evidence from oil producing economies in Africa. *Energy Strategy Reviews* **27**, 100434 (2020, 2020/01/01/). <https://doi.org/10.1016/j.esr.2019.100434>
- Bekun, F.V., Alola, A.A., Sarkodie, S.A.: Mar 20). Toward a sustainable environment: nexus between CO(2) emissions, resource rent, renewable and nonrenewable energy in 16-EU countries. *Sci Total Environ* **657**, 1023–1029 (2019). <https://doi.org/10.1016/j.scitotenv.2018.12.104>

- Dabboussi, M., Abid, M.: A comparative study of sectoral renewable energy consumption and GDP in the U.S.: Evidence from a threshold approach. *Renewable Energy* **192**, 705–715 (2022, 2022/06/01/). <https://doi.org/10.1016/j.renene.2022.03.057>
- Han, S., Lin, C., Zhang, B., Farnoosh, A.: Projections and recommendations for energy structure and industrial structure development in china through 2030: a system dynamics model. *Sustainability* **11**, 4901 (2019, 09/07). <https://doi.org/10.3390/su11184901>
- Nguyen, Hung, T., Thach, N.N.: A Panorama of Applied Mathematical Problems in Economics. *Thai Journal of Mathematics. Special Issue: Annual Meeting in Mathematics 1–20* (2018)
- Nguyen, H.T., Trung, N.D., Thach, N.N.: Beyond traditional probabilistic methods in econometrics. In: Kreinovich, V., Thach, N., Trung, N., Van Thanh, D. (eds) *Beyond Traditional Probabilistic Methods in Economics. ECONVN 2019. Studies in Computational Intelligence*, vol 809. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-04200-4\\_1](https://doi.org/10.1007/978-3-030-04200-4_1)
- Kreinovich, V., Thach, N.N., Trung, N.D., Thanh, D.V. (eds.): *Beyond Traditional Probabilistic Methods in Economics*. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-04200-4>
- Salari, M., Javid, R.J., Noghanibehambari, H.: The nexus between CO2 emissions, energy consumption, and economic growth in the U.S. *Economic Analysis and Policy* **69**, 182–194 (2021, 2021/03/01/). <https://doi.org/10.1016/j.eap.2020.12.007>
- Saldivia, M., Kristjanpoller, W., Olson, J.E.: Energy consumption and GDP revisited: a new panel data approach with wavelet decomposition. *Applied Energy* **272**, 115207 (2020, 2020/08/15/). <https://doi.org/10.1016/j.apenergy.2020.115207>
- Sasana, H., Ghozali, I.: International Journal of Energy Economics and Policy The Impact of Fossil and Renewable Energy Consumption on the Economic Growth in Brazil, Russia, India, China and South Africa. *Int. J. Ener. Econo. Poli.* **7**, 1–7 (2017)
- Stern, D.I.: A multivariate cointegration analysis of the role of energy in the US macroeconomy. *Energy Economics* **22**(2), 267–283 (2000, 2000/04/26/). [https://doi.org/10.1016/S0140-9883\(99\)00028-6](https://doi.org/10.1016/S0140-9883(99)00028-6)
- Thach, N.N.: How to Explain when the ES is Lower than One? A Bayesian Nonlinear Mixed-Effects Approach. *Journal of Risk and Financial Management* (2020). Retrieved February 20 from <<https://www.mdpi.com/1911-8074/13/2/21>>
- Tran, B.-L., Chen, C.-C., Tseng, W.-C.: Causality between energy consumption and economic growth in the presence of GDP threshold effect: Evidence from OECD countries. *Energy* **251**, 123902 (2022, 2022/07/15/). <https://doi.org/10.1016/j.energy.2022.123902>
- Yang, M., Wang, E.-Z., Hou, Y.: The relationship between manufacturing growth and CO2 emissions: Does renewable energy consumption matter? *Energy* **232**, 121032 (2021, 2021/10/01/). <https://doi.org/10.1016/j.energy.2021.121032>

# Author Index

## A

Autchariyapanitkul, Kittawit 169

## B

Baiya, Surapot 193  
Bokati, Laxman 169  
Briggs, William M. 38

## C

Canh, Tran Quang 427, 440  
Chaipunya, Parin 163  
Chaiwan, Anaspree 345, 358  
Charpentier, Arthur 45  
Chimprang, Namchok 637  
Chitkasame, Terdthiti 235  
Choy, S. T. Boris 107

## D

Dang, Nhan Truong Thanh 264, 498, 667  
Dang, Thuy T. 654  
Deepan, Urairat 163  
Diep, Nguyen Thi Ngoc 427, 440  
Dinh, Thi Thu Hong 588  
Dissanayake, Gnanadarsha Sanjaya 286

## E

Ersoy, Erkal 131

## F

Flachaire, Emmanuel 45

## G

Gallic, Ewen 45

## H

Ha, Doan Thanh 211  
Ha, Van Dung 264, 498  
Hallin, Marc 90  
Hoang, Dinh Cong 654  
Hoang, Tong Viet Bao 327  
Hoang, Tri Minh 533

Holguin, Sofia 174

Hue, Phan Thi Minh 403

Huyen, Nguyen Ngoc 248

## J

Jayathunga, Shanika Madushani 286

## K

Kaewsompong, Nachattapong 235  
Khoi, Bui Huy 299  
Klinlampu, Chaiwat 627  
Kosheleva, Olga 169, 178  
Kreinovich, Vladik 169, 174, 178, 186  
Kumam, Poom 163

## L

Lan, Le Thi 574  
Le, Thi Anh Tuyet 498, 667  
Le, Thi Thanh Huyen 449  
Leela-apiradee, Worrawate 544  
Leurcharusmee, Supanika 345, 358  
Li, Haoyang 131  
Lieu, Phan Thi 376, 691  
Linh, Nguyen Tran Xuan 654  
Lu, Shih-Hao 384

## M

Ma, Ziwei 107  
Maneejuk, Paravee 193  
My, Duong Tien Ha 403

## N

Nenna, Luca 117  
Ngoc, Bui Hoang 417  
Nguyen, Anh Tu 384  
Nguyen, Hau Trung 654  
Nguyen, Hung T. 1  
Nguyen, Phuong Hoang 174  
Nguyen, Thanh Phuc 588  
Nguyen, Tien Nhat 449  
Nguyen, Van Dan 667

Nguyen, Van Tung 264, 498, 667  
 Nhung, Pham Thi Hong 211

**P**

Pass, Brendan 117  
 Phaimekha, Sunisa 513  
 Pham, Nam Hai 523, 533  
 Pham, Uyen 178  
 Polard, Piangtawan 637  
 Poonsin, Tassathorn 544

**Q**

Quynh, Nguyen Thi Nhu 480

**R**

Rakpho, Pichayakone 235  
 Reyes, Christopher 186

**S**

Saijai, Worrawat 513, 560  
 Schaffer, Mark E. 131  
 Szendrei, Tibor 131

**T**

Tansuchat, Rongchai 627  
 Tarapituxwong, Supareuk 637  
 Thach, Nguyen Ngoc 299, 376, 403, 417,  
 427, 440, 465, 533, 574, 691  
 Thanabordeekij, Pithoon 193  
 Thanh, Bui Dan 211, 248, 465, 574

Thanh, Dang Nhan Truong 274  
 Thanh, Le Thi Phuong 327  
 Thanomsing, Vayu 544  
 Thao, Le Thi Phuong 327  
 Thongkairat, Sukrit 560  
 Thunjang, Thanakorn 544  
 Tran, Huong Thi Thanh 612  
 Tran, Ngoc Tho 588  
 Trung, Nguyen Duc 465, 480

**V**

Van Diep, Nguyen 403  
 Van Dung, Ha 274  
 Van Le, Chon 186, 315  
 Van Ngo, Tuan 533  
 Van Tung, Nguyen 274  
 Vo, Thuy Kieu Thi 523  
 Vu, Thuong Thi 315

**W**

Wang, Hung-Jen 384  
 Wang, Tonghui 107  
 Wei, Zheng 107

**Y**

Yamaka, Woraphon 150, 676

**Z**

Zhu, Xiaonan 107