

Contributions to Economics

M. Kenan Terzioğlu *Editor*

Advances in Econometrics,  
Operational Research,  
Data Science and  
Actuarial Studies

Techniques and Theories

 Springer

# **Contributions to Economics**

The series *Contributions to Economics* provides an outlet for innovative research in all areas of economics. Books published in the series are primarily monographs and multiple author works that present new research results on a clearly defined topic, but contributed volumes and conference proceedings are also considered. All books are published in print and ebook and disseminated and promoted globally.

The series and the volumes published in it are indexed by Scopus and ISI (selected volumes).

More information about this series at <https://link.springer.com/bookseries/1262>

M. Kenan Terzioğlu  
Editor

# Advances in Econometrics, Operational Research, Data Science and Actuarial Studies

Techniques and Theories

 Springer



*Editor*

M. Kenan Terzioğlu  
Faculty of Economics and Administrative  
Sciences  
Trakya University  
Edirne, Turkey

ISSN 1431-1933

ISSN 2197-7178 (electronic)

Contributions to Economics

ISBN 978-3-030-85253-5

ISBN 978-3-030-85254-2 (eBook)

<https://doi.org/10.1007/978-3-030-85254-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

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

Empirical/experimental economics and mathematics/statistics/econometrics are becoming more and more important as the data quality increases and the software performance improves. The effect of globalization and easy access to technology initiated the transformation to an information-oriented society. This transformation means complete and satisfactory statistical results, which can only be achieved through appropriate methods, are now a necessity. Additionally, emergence of specialized branches led to a movement away from the stereotypical patterns and development of new patterns that are more in line with the current economic and financial structure. Teaching the model structures might seem to be fundamental; however, gaining the ability to build and develop a model based on events/problems experienced in life is important both in terms of solving the problems and evaluating the new models/methods in the field of application. With this in mind, the models are established on real problems with the aim of developing new models and/or achieving policy proposals through appropriate analyses that would be beneficial in solving problems.

Processing, storage and communication of data became critical for all aspects of the global system as we entered the age of data. Data science emerged when digital existence expanded rapidly through digital transformation and evolved into global data, and it allows us uncover the data patterns through intuition. Analyses performed through models that are developed by uncovering the simplicity at the basis of complex processes and structures guides us by offering realistic solutions to problems. New approaches emerge as a result of the interaction between different disciplines as new algorithms are developed for analysis of data sequences. Within this structure, new disciplines stand out along with data mining, artificial intelligence, database management systems, and mathematical/statistical/econometric analysis techniques. As the way we address and scrutinize the problem changes; the mathematical-, statistical-, and econometrical-based system structures evolve as a result of the technological advancements. Increase of interest to data science field as a result of digital transformation is most prominent at universities and national/international R&D projects. Additionally, many individuals interested in the subject are trying to learn and apply data science practices using already existing

data sets. However, as the learning and practice steps, we mentioned above lack an end-to-end scenario, there is still missing information as to why data science solutions were developed and/or how they can be used. As artificial intelligence technologies are frequently used in different areas of Mathematics, Statistics and Econometrics; data science becomes a prominent ally to these fields. As the book defines problems based on an event/issue scenario, uses different data science approaches and also applies to other relevant scientific fields (econometrics, operations research and actuarial sciences), we aim to provide a comprehensive knowledge to the reader about the current theoretical and research studies in the relevant fields.

This book is designed to introduce the mathematical/statistical and econometrical underpinnings of the main tools used in empirical economics within in an effort to bridge the gap between analytic, closed-form methods, and numerical methods. Advanced methods and state-of-the-art models are used for forecasting and policy analysis in a wide range of applications in econometrics that quantifies causal relations among economic variables, operational research that solves optimization and decision-making problems, actuarial sciences that focuses on risk management to provide (support) maximum benefit (policy decisions) and data science that combines cleansing, mining, preparation, aligning, and analysis of the data. This book differs from other books as it tries to associate data science (data-scientists coming from different backgrounds) with some basic/advanced concepts and tools used in econometrics, operational research and actuarial sciences. Special emphasis will be given to four interrelated field: econometrics, operational research, actuarial sciences, and data science. Each chapter is created to gain a profound and detailed understanding of advanced theory and methods, to have knowledge about the design, estimation, forecasting and analysis of linear/nonlinear complex dynamic models, to raise awareness of evidence of asymptotic properties of estimators/test statistics and to understand the interplay between interrelated-field techniques and modeling assumptions both for advanced theoretical and practical applications that are used in the industries, bank/finance sectors, governments, think-tanks, and other sectors/institutes.

This book focuses on techniques and theories drawn from mathematics, statistics, computer science, and information science to analyze problems in business, economics, finance, insurance, and related fields. The book's research article format will be its most distinctive feature; it will offer the readers the opportunity to investigate several methods and approaches they can use in research from a wide perspective, and it will serve as a guide for new researches. When the scientific branches are analyzed individually, we can see that they are all associated and intertwined. Therefore, as opposed to other books, this book approaches econometrics, operations research and actuarial sciences together as a whole and lays out their association with data science. This book highlights solution proposals to clear the question marks in related fields using mathematical/statistical/actuarial modeling and data science by constructing/applying appropriate models with real-life data and designing/implementing computer algorithms to evaluate decision-making process. This would enhance the perspective of the researcher, building a basis for development of new/hybrid models. The chapters of the book will also offer examples

from the literature on the applied methods and approaches and explain the differences in comparison to the similar studies; thereby contributing to the enrichment and improvement of future research. The book also addresses the discussions on the deficiencies or benefits of the methods described in the book, which will offer additional value to the readers. This book, which is expected to be a comprehensive and beneficial reference by bringing together developments in different disciplines, also includes numerical samples and problem sets. This book serves as a reference for researches, academicians, graduate students, and related readers (including teachers and tutors) who have a background in a quantitative field such as mathematics, statistics, econometrics, economics, finance, computer science, actuarial science, insurance, and banking departments of the universities, as well as model designers developing programs/software for the business world. Those interested in the subject of the book can improve generic skills such as evaluation and synthesis of ideas/views/evidence, critical thinking, statistical reasoning, and problem-solving skills. Moreover, reader can have ability to evaluate current academics research published in journals and books, to reproduce existing related-field research, and to identify the key arguments and strategies underlying current and existing research.

Edirne, Turkey

Assoc. Prof. Dr. M. Kenan Terzioğlu

## Acknowledgements

I would like to express my gratitude to all academicians who have made valuable contributions and have provided invaluable assistance by taking part in the preparation of this book. I would like to thank the editorial staff of Springer for their invaluable support—Niko Chtouris, Mythili Settu, and Femina Joshi Arulthas. Special thanks to Prof. Dr. Serpil Terzioğlu and Dr. M. Nuri Terzioğlu for supporting me stepping into academia and Prof. Dr. Roger J. A. Laeven for raising my academic awareness in the academic development process. Thanks to Prof. Dr. Meral Sucu, Prof. Dr. Nezir Köse, Prof. Dr. Yeliz Yalçın, Prof. Dr. Nurcan Metin, Asst. Prof. Süreyya Temelli, and Asst. Prof. Yasemin Koldere Akın for their invaluable contribution to my academic development.

## About This Book

“Advances in Econometrics, Operations Research, Data Science and Actuarial Studies” based on article format in which the model/technique applications, results, and policy/decision suggestions stands out. In addition to understand the subject, it enables to develop different or hybrid models by improving the perspective of the researcher. This book is designed to analyze complex issues and to make recommendations in economics, business, and finance by combining mathematics, statistics, econometrics, actuarial sciences, and data science branches. It improves the perspective by providing information about the comprehensive applications and methods. It provides opinions in differentiating the application areas of models or in creating new and hybrid model structures by increasing the ability to synthesize and interpret.

*Roman G. Smirnov, Kunpeng Wang, and Ziwei Wang* examine the growth of the Chinese economy since the opening of China in 1978. Specifically, the authors employ statistical and mathematical methods to investigate the data for the period in terms of its compatibility with the Cobb–Douglas production function.

*Mehmet Özcan and Funda Yurdakul* develop a new nonlinear unit root test process that models a structural break and a regime switching together. It employs a more complex threshold autoregressive model to investigate unit roots which was introduced by Caner and Hansen (2001). The simulation findings of this chapter confirm that the new approach is a useful tool to test the presence of a unit root for nonlinear time series.

*Emawtee Bissoondoyal-Bheenick, Robert Brooks, and Hung Xuan Do* provide a comprehensive study to answer whether jumps are transmitted across the European foreign exchange market and, if yes, to what extent they are transmitted. This chapter reveals the static and dynamic behavior of the jump connectedness (spillover) among five G10 European currencies.

*Emel Kızılok Kara, Sibel Açık Kemaloğlu, and Ömer Ozan Evkaya* examine the potential benefits of asymmetric copula modeling and explain the dependence among currencies using Khoudraji copula models. This chapter is focused on explaining the direction of the dependency among the currencies with the considered models.

*Kadir Y. Eryiğit and Veli Durmuşoğlu* present the set of international parity conditions and joint tests of the validity of uncovered interest parity (UIP)

and purchasing power parity (PPP) focusing on possible structural breaks such as the global financial crisis and implementation of macroprudential policies after the global financial crisis.

*Akram S. Hasanov and Salokhiddin S. Avazkhodjaev* evaluate the need to incorporate the information on structural breaks to improve the forecast accuracy, investigate whether stochastic volatility models' forecasting performance is improved than GARCH-type models when the endogenously detected breaks are considered and analyzed the role of distributions in volatility prediction performance. This chapter examines the forecasting performance of stochastic volatility models by incorporating breaks detected through the ICSS algorithm developed by Sansó et al. (2004).

*Hakan Eygü* uses statistical quality control charts to monitor small shifts in the process mean. This chapter controls charts, which is one of the statistical process control methods, performance measurement, and compares sampling methods. The results show that the sampling plan to be applied is a function of the magnitude of the process shift.

*Yasin Büyükkör and A. Kemal Şehirlioğlu* propose a new M-estimation method for asymmetrical and heavy long-tailed data set based on the Pearson Type IV distribution. When the anomalies which are asymmetry, excess kurtosis and heavy-long tailness occur in data, traditional M-estimators (Least Squares, Huber M-Estimator, and Tukey M-Estimator) can not achieve good solution in regression analysis. Therefore, in this chapter, regression parameters is estimated by using weight function of proposed M-estimation method when data follows PIV.

*Annah Gosebo, Donald Makhalemele, and Zinhle Simelane* examine hedging strategies of South African real estate investment trusts (REITs) using discrete volatility models. The significance of exploring the differences of the dynamics that exist between the REITs and common stocks aids in understanding the returns and risks movements.

*Avni Önder Hanedar* investigates those political events that could create substantial price fluctuations in commodity markets through increasing uncertainty. The chapter fills the gap on modeling difficulty to set causal estimate for the effects of political uncertainty on good prices.

*Ebru Çağlayan-Akay and Zamira Oskonbaeva* investigate whether foreign direct investment causes environmental pollution or not and which hypothesis is valid in selected transition economies. The chapter provides empirical evidence supporting the pollution halo hypothesis by highlighting the existence of long-run asymmetric cointegration.

*Fela Özbey* examines, in light of the economics literature emphasizing the prevalence of nonlinearity and asymmetry in the behavior of economic agents, the asymmetries of ERPT because of the downward stickiness of prices, and the cruciality of pre-testing for explosive and seasonal roots due to the limitations of the bounds testing procedure steps that are often neglected in empirical studies.

*Eda Yalçın Kayacan and Aygül Anavatan* investigate whether the macroeconomic factors affect the URAP global ranking measuring academic performance. The aim of the chapter is to identify country-specific factors that are thought to influence

academic performance and to reveal how the country-specific factors should be developed for universities to increase their success.

*Ferda Esin Gülel* and *Öncü Yanmaz Arpacı* aim to measure the impact of outsourcing and innovation on Industry 4.0 in Fortune 500 companies in Turkey. The significance of outsourcing, satisfaction in outsourcing, and innovation factors are examined for industries switching to Industry 4.0.

*Süreyya Temelli* and *Mustafa Seviüktekin* provide, with a limited number of quantitative studies in order to use microdata for macro-level policymaking, evidence for social politicians by empirically investigating the impact of social assistance on subjective well-being.

*María Francisca Peñaloza Talavera*, *Jaime Apolinar Martínez-Arroyo*, and *Marco Alberto Valenzo-Jiménez* analyze the viability of the emergence of a fishing and aquaculture cluster that promotes the regional competitiveness of the territory and determine the feasibility of the formation of an agglomeration of companies using the methodology proposed by Fregoso (2012)

*Mine Aydemir* and *Nuran Bayram Arlı* aim to put forward the effects of personality types and self-efficacy on deep and surface learning.

*Radka Nacheva* proposes a research framework for evaluation of emotional attitudes in social media and test emotions mining research framework.

*Pejman Peykani*, *Jafar Gheidar-Kheljani*, *Donya Rahmani*, *Mohammad Hossein Karimi Gavareshki*, and *Armin Jabbarzadeh* propose a novel uncertain super-efficiency data envelopment analysis (USEDEA) approach to the ranking of DMUs in the presence of uncertain data. The chapter develops a new method for ranking homogeneous decision-making units in the presence of uncertain inputs and/or outputs.

*İpek Deveci Kocakoç* and *Gökçe Baysal Türkölmez* aim to design an uncrowded hospital by using data mining analysis of patient data for minimization the spending period in the hospitals and the intense circulation between the clinics reducing the risk of transmissio, especially in the COVID-19 outbreak and other seasonal epidemics.

*Nikos Chatzistamoulou* and *Phoebe Koundouri* adopt a non-parametric metafrontier framework to handle technological heterogeneity and calculate the productive performance and environmental efficiency through the data envelopment analysis (DEA) and directional distance function (DDF)

*Ezgi Demir*, *Rana Elif Dinçer*, *Batuhan Atasoy*, and *Sait Erdal Dinçer* focus on personnel selection in the field of human resources management of artificial intelligence. Although the use of artificial intelligence in human resources management is limited, this chapter sets an example with the algorithms and methodology used. In this way, artificial intelligence algorithms have been developed in the field of human resources.

*Pejman Peykani*, *Mohammad Namakshenas*, *Mojtaba Nouri*, *Neda Kavand* and, *Mohsen Rostamy-Malkhalifeh* propose the possibilistic portfolio optimization (PPO) model for PO problem under uncertainty and ambiguity. The chapter develops a new fuzzy portfolio optimization model that is capable to be used in the presence of fuzzy data and linguistic variables.



*Nilsen Kundakcı* aims to evaluate energy service company (ESCO) selection process of a textile firm. The main contribution of this chapter is the integrated use of fuzzy SWARA and fuzzy MARCOS methods, which are two relatively new fuzzy multi-criteria decision-making methods. The proposed approach provides a new insight into the selection process of ESCOs for the firms. Since the developed approach shows a great deal of flexibility, it can be used in other real-life decision problems of firms.

*Ioanna Papadaki* and *Michail Tsagris* examine the relationship NBA players' performance on court and their salaries with the ultimate aim to predict the players' salaries using their game statistics, to discriminate between low- and high-paid players and, more like a by-product, to detect of over-paid players based on advanced statistical and machine learning algorithms and techniques

*Seyyide Doğan*, *Deniz Koçak*, and *Murat Atan* aim to provide how success or failure situations of firms, which are trade-in Borsa İstanbul by means of support vector machine. This chapter indicates that feature selection phase is almost important to obtain a good prediction model for financial problems.

*Darya Lapitskaya*, *Hakan Eratalay*, and *Rajesh Sharma* offer an empirical comparison of the financial econometrics and machine and deep learning methods in predicting stock returns using S&P 100 historical returns and COVID-19-related news sentiments. This chapter focuses on the predictive performance of these methods both for in-sample and out-of-sample predictions.

*Olgun Aydin* and *Krystian Zieliński* analyze the residential market in Trójmiasto using clustering techniques while offering detailed insights for the secondary and primary residential markets separately.

*Michail Tsagris* and *Stefanos Fafalios* aim to detect the most important characteristics and evaluate their effect on the car's price for purchasing a value-for-money car and being aware of the factors that affect the price using nonlinear machine learning models are employed revealing interesting patterns.

*Cem Yavrum* and *A. Sevtap Selcuk-Kestel* investigate the influence of outliers in mortality rates on the annuity prices as such outliers in mortality rates can have significant volatility on mortality estimation. The main focus of this chapter is the implementation of the outlier-adjusted Lee–Carter model to mortality rates at which possible outliers are detected and removed simultaneously with an iteration process.

*Jinhui Zhang*, *Sachi Purcal*, and *Jiaqin Wei* investigate the optimal behavior of an agent with a bequest motive in volatile circumstances—in financial environment with price jumps and various switching (parameter) regimes. This chapter develops an optimal strategy by extending a model formulated by Scott Richard in 1975 which includes insurance motives, itself an extension of the seminal stochastic control work of Robert Merton in 1969 for the case of an individual investor and sets out to determine retirees' optimal consumption, investment, and insurance decisions.

*Ashlhan Şentürk Acar* aims to compare the predictive performance of logistic regression and ensemble machine learning methods for the prediction of claim probability in the presence of excess zeros. This chapter examines the predictive performance of logistic regression and ensemble methods when the dependent variable is binary with highly imbalanced data.

*Amela Omerašević* and *Jasmina Selimović* explore and analyze the benefits of risk classification methods by using the data mining techniques on premium ratemaking in nonlife insurance. The improvement in the process of nonlife insurance premium ratemaking is reflected in the choice of predictors or risk factors that have an impact on insurance premium rates.

*Željko Šain*, *Edin Taso*, and *Jasmina Selimović* actualize the application of modern methods in the management of the investment portfolio of a (re) insurance company, primarily methods for managing the risks of the securities portfolio and methods for managing credit risk. The chapter focuses the current domestic regulations on insurance, as well as the prescribed methodology for assessing the main risks and for quantitative analysis of the impact of the investment portfolio on the solvency of (re) insurance companies in BiH.

# Contents

<b>The Cobb-Douglas Production Function for an Exponential Model</b> . . . . .	1
Roman G. Smirnov, Kunpeng Wang, and Ziwei Wang	
<b>Threshold Unit Root Tests with Smooth Transitions</b> . . . . .	13
Mehmet Özcan and Funda Yurdakul	
<b>Jump Connectedness in the European Foreign Exchange Market</b> . . . . .	31
Emawtee Bissoondoyal-Bheenick, Robert Brooks, and Hung Xuan Do	
<b>Modeling Currency Exchange Data with Asymmetric Copula Functions</b> . . . . .	49
Emel Kızılok Kara, Sibel Açık Kemaloğlu, and Ömer Ozan Evkaya	
<b>The Joint Tests of the Parity Conditions: Evidence from a Small Open Economy</b> . . . . .	63
Kadir Y. Eryiğit and Veli Durmuşoğlu	
<b>Stochastic Volatility Models with Endogenous Breaks in Volatility Forecasting</b> . . . . .	81
Akram S. Hasanov and Salokhiddin S. Avazkhodjaev	
<b>Effect in Quality Control Based of Hotelling <math>T^2</math> and CUSUM Control Chart</b> . . . . .	99
Hakan Eygü	
<b>A Robust Regression Method Based on Pearson Type VI Distribution</b> . . . . .	117
Yasin Büyükkör and A. Kemal Şehirlioğlu	
<b>Discrete Volatilities of Listed Real Estate Funds</b> . . . . .	143
Annah Gosebo, Donald Makhalemele, and Zinhle Simelane	
<b>Have Commodity Markets Political Nature?</b> . . . . .	171
Avni Önder Hanedar	

<b>A Nonlinear Panel ARDL Analysis of Pollution Haven/Halo Hypothesis</b> .....	189
Ebru Çağlayan-Akay and Zamira Oskonbaeva	
<b>An Investigation of Asymmetries in Exchange Rate Pass-Through to Domestic Prices</b> .....	207
Fela Özbey	
<b>Investigation of the Country-Specific Factors for URAP</b> .....	221
Eda Yalçın Kayacan and Aygül Anavatan	
<b>The Impact of Outsourcing and Innovation on Industry 4.0</b> .....	235
Ferda Esin Gülel and Öncü Yanmaz Arpacı	
<b>Subjective Well-Being of Poor Households</b> .....	251
Süreyya Temelli and Mustafa Sevüktekin	
<b>Formation of a Fishing and Aquaculture Cluster as a Tool for Regional Competitiveness</b> .....	267
María Francisca Peñaloza-Talavera, Jaime Apolinar Martínez-Arroyo, and Marco Alberto Valenzo-Jiménez	
<b>A Path Analysis of Learning Approaches, Personality Types and Self-Efficacy</b> .....	285
Mine Aydemir and Nuran Bayram Arlı	
<b>Emotions Mining Research Framework: Higher Education in the Pandemic Context</b> .....	299
Radka Nacheva	
<b>Uncertain Super-Efficiency Data Envelopment Analysis</b> .....	311
Pejman Peykani, Jafar Gheidar-Kheljani, Donya Rahmani, Mohammad Hossein Karimi Gavareshki, and Armin Jabbarzadeh	
<b>Using Data Mining Techniques for Designing Patient-Friendly Hospitals</b> .....	321
İpek Deveci Kocakoç and Gökçe Baysal Türkölmez	
<b>Sustainability Transition Through Awareness to Promote Environmental Efficiency</b> .....	345
Nikos Chatzistamoulou and Phoebe Koundouri	
<b>Data Mining Approach in Personnel Selection: The Case of the IT Department</b> .....	363
Ezgi Demir, Rana Elif Dinçer, Batuhan Atasoy, and Sait Erdal Dinçer	
<b>A Possibilistic Programming Approach to Portfolio Optimization Problem Under Fuzzy Data</b> .....	377
Pejman Peykani, Mohammad Namakshenas, Mojtaba Nouri, Neda Kavand, and Mohsen Rostamy-Malkhalifeh	

**A Hybrid Fuzzy MCDM Approach for ESCO Selection** ..... 389  
 Nilsen Kundakci

**Are NBA Players’ Salaries in Accordance with Their Performance on Court?** ..... 405  
 Ioanna Papadaki and Michail Tsagris

**Financial Distress Prediction Using Support Vector Machines and Logistic Regression** ..... 429  
 Seyyide Doğan, Deniz Koçak, and Murat Atan

**Predicting Stock Returns: ARMAX versus Machine Learning** ..... 453  
 Darya Lapitskaya, Hakan Eratalay, and Rajesh Sharma

**Analysing the Residential Market Using Self-Organizing Map** ..... 465  
 Olgun Aydin and Krystian Zieliński

**Advanced Car Price Modelling and Prediction** ..... 479  
 Michail Tsagris and Stefanos Fafalios

**Impact of Outlier-Adjusted Lee–Carter Model on the Valuation of Life Annuities** ..... 495  
 Cem Yavrum and A. Sevtap Selcuk-Kestel

**Optimal Life Insurance and Annuity Demand with Jump Diffusion and Regime Switching** ..... 515  
 Jinhui Zhang, Sachi Purcal, and Jiaqin Wei

**Prediction of Claim Probability with Excess Zeros** ..... 531  
 Aslıhan Şentürk Acar

**Risk Classification in Nonlife Insurance Premium Ratemaking** ..... 541  
 Amela Omerašević and Jasmina Selimović

**Insurance Investments Management—An Example of a Small Transition Country** ..... 573  
 Željko Šain, Edin Taso, and Jasmina Selimović

# Contributors

**Ashhan Şentürk Acar** Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey

**Sibel Açık Kemaloğlu** Department of Statistics, Faculty of Sciences, Ankara University, Ankara, Turkey

**Aygül Anavatan** Department of Econometrics, Faculty of Economics and Administrative Sciences, Pamukkale University, Denizli, Turkey

**Öncü Yanmaz Arpacı** Pamukkale University, Denizli, Turkey

**Murat Atan** Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey

**Batuhan Atasoy** Department of Mechanical Engineering, Piri Reis University, Tuzla, Turkey

**Salokhiddin S. Avazkhodjaev** Tashkent Institute of Finance, Tashkent, Republic of Uzbekistan

**Mine Aydemir** Faculty of Economics and Administrative Sciences, Bursa Uludağ University, Bursa, Turkey

**Olgun Aydın** Gdańsk University of Technology, Gdansk, Poland

**Nuran Bayram Arlı** Faculty of Economics and Administrative Sciences, Bursa Uludağ University, Bursa, Turkey

**Emawtee Bissoondoyal-Bheenick** School of Economics, Finance and Marketing, RMIT University, Melbourne, Australia

**Robert Brooks** Department of Econometrics and Business Statistics, Monash University, Clayton, Australia

**Yasin Büyükkör** Faculty of Economics and Administrative Sciences, Department of Econometrics, Karamanoğlu Mehmetbey University, Karaman, Turkey

**Ebru Çağlayan-Akay** Department of Econometrics, Marmara University, Istanbul, Turkey

**Nikos Chatzistamoulou** School of Economics and Research Laboratory On Socio-Economic and Environmental Sustainability–ReSEES, Athens University of Economics and Business, Athens, Greece;  
Department of Economics, University of Ioannina, Ioannina, Greece

**Ezgi Demir** Department of Management Information Systems, Piri Reis University, Tuzla, Turkey

**Rana Elif Dinçer** Department of Business Informatics, Marmara University, Göztepe, Turkey

**Sait Erdal Dinçer** Department of Econometrics, Marmara University, Göztepe, Turkey

**Hung Xuan Do** School of Economics and Finance, Massey University, Palmerston North, New Zealand

**Seyyide Doğan** Department of Econometrics, Karamanoğlu Mehmetbey University, Karaman, Turkey

**Veli Durmuşoğlu** Department of Econometrics, Bursa Uludağ University, Bursa, Turkey

**Hakan Eratalay** School of Economics and Business Administration, University of Tartu, Tartu, Estonia

**Kadir Y. Eryiğit** Department of Econometrics, Bursa Uludağ University, Bursa, Turkey

**Ömer Ozan Evkaya** Department of Statistical Sciences, Università Di Padova, Padova, Italy

**Hakan Eygü** Faculty of Economics and Administrative Sciences, Department of Econometrics, Statistical Research, Ataturk University, Erzurum, Turkey

**Stefanos Fafalios** Gnosis Data Analysis, Herakleion, Crete, Greece

**Jafar Gheidar-Kheljani** Management and Industrial Engineering Department, Malek Ashtar University of Technology, Tehran, Iran

**Annah Gosebo** School of Construction Economics & Management, WITS University, Johannesburg, South Africa

**Ferda Esin Gülel** Pamukkale University, Denizli, Turkey

**Avni Önder Hanedar** Faculty of Political Sciences, Sakarya University, Serdivan, Turkey

**Akram S. Hasanov** Monash University Malaysia, Kuala Lumpur, Selangor, Malaysia

**Armin Jabbarzadeh** Department of Automated Production Engineering, École de Technologie Supérieure (ETS), Montreal, Canada

**Mohammad Hossein Karimi Gavareshki** Management and Industrial Engineering Department, Malek Ashtar University of Technology, Tehran, Iran

**Neda Kavand** Department of Mathematics, Faculty of Basic Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran

**Eda Yalçın Kayacan** Department of Statistics, Faculty of Science and Literature, Pamukkale University, Denizli, Turkey

**Emel Kızılok Kara** Department of Actuarial Sciences, Faculty of Arts and Sciences, Kırıkkale University, Kırıkkale, Turkey

**Deniz Koçak** Department of Econometrics, Osmaniye Korkut Ata University, Osmaniye, Turkey

**İpek Deveci Kocakoç** Econometrics Department, Dokuz Eylül University, Izmir, Turkey

**Phoebe Koundouri** School of Economics and Research Laboratory On Socio-Economic and Environmental Sustainability–ReSEES, Athens University of Economics and Business, Athens, Greece;

Director, Sustainable Development Unit, ATHENA Research Center, Co-Chair, UN SDSN Europe Fellow, Academy of Art and Science, Marousi, Greece

**Nilsen Kundakcı** Department of Business Administration, Pamukkale University, Denizli, Turkey

**Darya Lapitskaya** School of Economics and Business Administration, University of Tartu, Tartu, Estonia

**Donald Makhalemele** School of Construction Economics & Management, WITS University, Johannesburg, South Africa

**Jaime Apolinar Martínez-Arroyo** Faculty of Accounting and Administrative Sciences, The Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico

**Radka Nacheva** University of Economics – Varna, Varna, Bulgaria

**Mohammad Namakshenas** School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

**Mojtaba Nouri** School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

**Amela Omerašević** Uniqa osiguranje d.d, Sarajevo, Bosnia and Herzegovina

**Zamira Oskonbaeva** Department of Economics, Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan



**Fela Özbey** Department of Econometrics, Faculty of Economics and Administrative Sciences, Çukurova University, Adana, Turkey

**Mehmet Özcan** Department of Economics, Faculty of Economics and Administrative Sciences, Karamanoglu Mehmetbey University, Karaman, Turkey

**Ioanna Papadaki** Department of Economics, University of Crete, Heraklion, Greece

**María Francisca Peñaloza-Talavera** Faculty of Accounting and Administrative Sciences, The Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico

**Pejman Peykani** School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

**Sachi Purcal** Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, China

**Donya Rahmani** Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran

**Mohsen Rostamy-Malkhalifeh** Department of Mathematics, Faculty of Basic Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran

**Željko Šain** School of Economics and Business, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

**A. Kemal Şehirlioğlu** Faculty of Economics and Administrative Sciences, Department of Econometrics, Dokuz Eylül University, İzmir, Turkey

**A. Sevtap Selcuk-Kestel** Institute of Applied Mathematics, Actuarial Sciences, Middle East Technical University, Ankara, Turkey

**Jasmina Selimović** School of Economics and Business, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

**Mustafa Sevüktekin** Department of Econometrics, Uludağ University, Bursa, Turkey

**Rajesh Sharma** Institute of Computer Science, University of Tartu, Tartu, Estonia

**Zinhle Simelane** School of Construction Economics & Management, WITS University, Johannesburg, South Africa

**Roman G. Smirnov** Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

**Edin Taso** Insurance Supervisory Agency in FBiH, Sarajevo, Bosnia and Herzegovina

**Süreyya Temelli** Department of Econometrics, Trakya University, Edirne, Turkey

**Michail Tsagris** Department of Economics, University of Crete, Heraklion, Greece;  
Department of Economics, University of Crete, Gallos Campus, Rethymnon, Crete, Greece

**Gökçe Baysal Türkölmez** Econometrics Department, Dokuz Eylül University, Izmir, Turkey

**Marco Alberto Valenzo-Jiménez** Faculty of Accounting and Administrative Sciences, The Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico

**Kunpeng Wang** Sichuan University-Pittsburgh Institute (SCUPI), Sichuan University, Chengdu, Sichuan, China

**Ziwei Wang** Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

**Jiaqin Wei** Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, China

**Cem Yavrum** Institute of Applied Mathematics, Actuarial Sciences, Middle East Technical University, Ankara, Turkey

**Funda Yurdakul** Department of Econometrics, Faculty of Economics and Administrative Sciences, Ankara Hacı Bayram Veli University, Ankara, Turkey

**Jinhui Zhang** Department of Actuarial Studies and Business Analytics, Macquarie Business School, Macquarie University, Sydney, Australia

**Krystian Zieliński** PwC Polska, Gdańsk, Poland

# Abbreviations

AC	Absorptive Capacity
AD	Absolute Deviation
ADABoost	Adaptive Boosting
ADF	Augmented Dickey Fuller
AGFI	Adjusted Goodness of Fit
AHP	Analytic Hierarchy Process
AIC	Akaike Information Criterion
AICC	Corrected AIC
AMG	Augmented Mean Group
AMOS	Analysis of Moment Structures
ANN	Artificial Neural Networks
AO	Additive Outlier
APARCH	Asymmetric Power Autoregressive Conditional Heteroskedasticity
API	Application Programming Interface
AR	Autoregressive
ARAS	Additive Ratio Assessment
ARCH	Autoregressive Conditional Heteroskedasticity
ARE	Asymptotic Relative Efficiency
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ARMAX	Autoregressive Moving Average Model with Exogenous Variables
ARWU	Academic Ranking of World Universities
ASEAN	The Association of Southeast Asian Nations
AUD	Australian Dollar
BAS	Bid–Ask Spread
BCC	Banker–Charnes–Cooper
BFGS	Broyden, Fletcher, Goldfarb and Shanno
BFI	Big Five Inventory
BIC	Bayesian Information Criterion
BiH	Bosnia and Herzegovina
BIST	Borsa Istanbul

BNS-G	Barndoff–Nielsen and Shephard G Jump Statistic
C12	Independent-Clayton Asymmetric Copula Model
C13	Independent-Frank Asymmetric Copula Model
C14	Independent-Gumbel Asymmetric Copula Model
CAD	Canadian Dollar
CADF	Cross-sectional Augmented Dickey Fuller
CAIC	Consistent AIC
CART	Classification and Regression Trees
CC	Cluster Coefficient
CCEMG	Common Correlated Effects Mean Group
CCP	Chance-Constrained Programming
CCR	Charnes-Cooper-Rhodes
CD test	Cross Section Dependence Test
CFI	Comparative Fit Index
CHAID	Chi-square Automatic Interaction Detector
CHE	Centre for Higher Education Development
CHEER	Capital-enhanced Equilibrium Exchange Rate
CHF	Swiss Franc
CIP	Covered Interest Rate Parity
CIPS	Cross-Sectional Augmented Im Pesaran Shin
CIS	Community Innovation Survey
CML	Capital Market Law
CNN	Convolutional Neural Networks
CNY	Chinese Yuan Currency
CO <sub>2</sub>	Carbon Dioxide
COPRAS	Complex Proportional Assessment
CPI	Consumer Price Index
CRISP-DM	Cross-Industry Standard Process Model for Data Mining
CRRA	Constant Relative Risk-Aversion
CRS	Constant Returns to Scale
CSY	China's Statistical Yearbook
CULI	Coefficient of Economic Unit per Labor in the Industry
CULS	Coefficient of Economic Unit per Work in the Sector
CUSUM	Cumulative Sum
CUSUMSQ	Cumulative Sum of Squares
CWTS	Centrum voor Wetenschap en Technologische Studies
CWUR	Center for World University Rankings
DA	Discriminant Analysis
DDF	Directional Distance Function
DEA	Data Envelopment Analysis
DGP	Data Generation Process
DMM	Data Mining Method
DMU	Decision Making Unit
DOLS	Dynamic Ordinary Least Squares
DPO	Deterministic Portfolio Optimization

DT	Decision Trees
E	Energy Use
EA/EAIR	Environmentally Aware
ECB	European Central Bank
ECO	Wilder Hill Clean Energy Index
ECR	Error Correction Representation
EGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity
EGD	European Green Deal
EnvEff	Environmental Efficiency
ERIX	European Renewable Energy Index
ERPT	Exchange Rate Pass-Through
ES	Exponentially Smoothing
ESCO	Energy Service Company
EU	European Union
EUR	Euro Currency
FA	Factor Analysis
FDI	Foreign Direct Investment
FDP	Financial Distress Prediction
FE	Fixed Effect
FGM	Farlie Gumbel Morgenstern
FN	False Negative
FO	Fuzzy Optimization
FP	False Positive
FPP	Fuzzy Mathematical Programming
FRM	Fractional Regression Model
FUCOM	Full Consistency Method
FX	Foreign Exchange
GA	Genetic Algorithm
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GBP	British Pound
GCC	The Gulf Cooperation Council
GCI	Global Competitiveness Index
GDP	Gross Domestic Product
GFC	Global Financial Crisis
GFEVD	Generalized Forecast-Error Variance Decomposition
GFI	Goodness of Fit
GGDC	Groningen Growth and Development Centre
GHYP	Generalized Hyperbolic Distribution
GJR	Glosten, Jagannathan and Runkle
GLM	Generalized Linear Model
GMM	Generalized Method of Moments
GOF	Goodness of Fit
GP	Genetic Programming
GSE	General Self-Efficacy Scale

GSPRE	Generalized Spatial Panel Random Effects Model
HDI	Human Development Index
HEEACT	Higher Education Evaluation and Accreditation Council of Taiwan
HEGY	Hylleberg, Engle, Granger and Yoo
HJB	Hamilton–Jacobi–Bellman
HM	Historical Mean
HMAE	Heteroskedasticity (adjusted) Mean Absolute Error
HMSE	Heteroskedasticity (adjusted) Mean Squared Error
HTML	Hypertext Markup Language
ICA	Independent Components Analysis
ICSS	Iterative Cumulative Sum Of Squares
IF	Influence Function
IFE	International Fisher Effect
IGARCH	Integrated Generalized Autoregressive Conditional Heteroskedasticity
IMF	International Monetary Fund
INEGI	National Institute of Statistics and Geography
IO	Innovational Outlier
I-O	Input-Oriented
IP	Interest Rate Parity
IRWLS	Iteratively Re-Weighted Least Squares
IT	Information Technologies
JB test	Jarque–Bera Test
JPY	Japanese Yen
JSE	Johannesburg Stock Exchange
K	Capital Stock
KKT	Karush–Khun–Tucker
KNN	K-nearest Neighbors Algorithm
k-NN	k-nearest Neighbour
Kt	Kilo Tones
L	Labor—Persons Engaged
LAD	Least Absolute Deviation
LC	Lee Carter
LEA/LEAIR	Less Environmentally Aware
LL	Log-Likelihood
LLC Test	Levin Lin Chu Test
LM	Lagrange Multiplier
LNV	Leybourne, Newbold and Vougas
LR	Likelihood Ratio
LRA	Logistic Regression Analysis
LS	Level Shift
LS	Least Squares
LSTM	Long Short-term Memory Algorithm
LTS	Least Trimmed Squares
LVQ	Learning Vector Quantization

MACBETH	Measuring Attractiveness through a Categorical-Based Evaluation Technique
MAD	Median Absolute Deviation
MADB	Mean-Absolute Deviation-Beta
MAE	Mean Absolute Error
MARCOS	Measurement Alternatives and Ranking according to the Compromise Solution
MARS	Multivariate Adaptive Regression Splines
Max.	Maximum
MCC	Matthews Correlation Coefficient
MCDM	Multi-Criteria Decision-Making
MCS	Model Confidence Set
MDA	Multi-dimensional Analysis
MF	Meta Frontier
MG	Mean Group
MIKTA	Mexico, Indonesia, the Republic of Korea, Turkey, and Australia
Min.	Minimum
MINT	Mexico, Indonesia, Nigeria, and Turkey
ML	Maximum Likelihood
ML	Machine Learning
MLP	Multi-layer Perceptron
MML	Metafrontier Malmquist–Luenberger Index
MoM	Method of Moments
MPL	Maximum Pseudo-Likelihood
MSE	Mean Squared Error
MSMEs	Micro, Small and Medium-Sized Enterprises
MTR	Meta-Technology Ratio
MVDA	Multiple Variable Discriminant Analysis
N	Standard Normal
NARDL	Nonlinear Autoregressive Distributed Lag
NAV	Net Asset Value
NB	Naïve Bayes
NFI	Normed Fit Index
NLP	Neuro-linguistic Programming
NLS	Nonlinear Least Square
NNFI	Non-Normed Fit Index
NOK	Norwegian Krone
NTU	National Taiwan University Ranking
OALC	Outlier-Adjusted Lee Carter
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary Least Squares
ONS	British Office for National Statistics
OO	Output-Oriented
PCA	Principal Components Analysis
PDE	Pearson Differential Equation

PDF	Pearson Distribution Family
pdf	Probability Density Function
PDS	Pearson Distribution System
PI	Pearson Type I Distribution
PIV	Pearson Type IV Distribution
PLS	Property Loan Stocks
PMADB	Possibilistic Mean-Absolute Deviation-Beta
PMG	Pooled Mean Group
PO	Portfolio Optimization
POMS	Profile of Mood States
PPE	Personal Protective Equipment
PPO	Possibilistic Portfolio Optimization
PPP	Purchasing Power Parity
PPS	Production Possibility Set
ProdPerf	Productive Performance
PSS	Pesaran, Shin, and Smith
PTE	Value of the Working-Age Population in the Region
PUT	Property Unit Trusts
PVI	Pearson Type VI Distribution
QLIKE	Quasi-likelihood
QQ	Quantile-to-quantile Plot
QS	Quacquarelli Symonds
R&D	Research and Development
RBF	Radial Basis Function
Rdbms	Relational Database Management System
Rec	Renewable Energy Consumption
REIT	Real Estate Investment Trust
RE-SEM	Random Effect with Spatial Error Model
RMR	Root Mean Square Residual
RMSE	Root Mean Squared Error
RMSEA	Root Mean Square Error of Approximation
RNN	Recurrent Neural Networks
RO	Robust Optimization
RSF	Reporters Without Borders
R-SPQ-2F	The Revised Two Factor Study Process Questionnaire
Rtrn	True Return Series
RUB	Russian Ruble currency
S&P	Standard & Poor's
S&P GCE	S&P Global Clean Energy Index
S&P100	Standard and Poor's 100 index
S.E.	Standard Error
SAC	Spatial Autoregressive Model with Autoregressive Disturbances
SAL	Student Approaches to Learning
SAP	System, Application & Products
SAR	Spatial Autoregressive Model



SARAR	Spatial Autoregressive Lag and Error Model
SARS	Severe Acute Respiratory Syndrome
SCIAN	North American Industry Classification System 2018
SDGs	Sustainable Development Goals
SDM	Spatial Durbin Model
SEDEA	Super-Efficiency Data Envelopment Analysis
SEK	Swedish Krona
SEM	Spatial Error Model
SEM-RE	Spatial Error Model with Random Effect
SJTUIHE	Shanghai Jiao Tong University Institute of Higher Education
SMEs	Small and Medium-sized Enterprises
SO	Stochastic Optimization
SPQ	Study Process Questionnaire
SPSS	Statistical Package of Social Science
SQP	Sequential Quadratic Programming
SRMR	Standardized Root Mean Square Residual
SSM	Superior Set of Models
SSTD	Skew Standardized Student
St.D.	Standard Deviation
STD	Standardized Student
STIRPAT	Stochastic (ST) Estimation of Environmental Impacts (I) by Regression (R) on Population (P), Affluence (A) and Technology (T)
SV	Stochastic Volatility
SVM	Support Vector Machine
SWARA	Stepwise Weight Assessment Ratio Analysis
TAB	Total Absolute Deviation
TAR	Threshold Autoregressive
TARCH	Threshold Autoregressive Conditional Heteroskedasticity
TC	Tehran Stock Exchange
Tg	Technology Gap
THE	Temporary Change
TN	True Negative
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TP	True Positive
TRY	Turkish Liras Currency
TSE	Times Higher Education
TTI	Total Labor Value in the Industry.
TTS	Total Labor Value in the Sector.
TUBITAK	Scientific and Technological Research Council of Turkey
UCC	Uncertain Chance Constraint
UDEA	Uncertain Data Envelopment Analysis
UEI	Value of Economic Units in the Industry.
UES	Value of Economic Units in the Sector.
UIP	Uncovered Interest Rate Parity
UK	United Kingdom

UNICEF	United Nations International Children's Emergency Fund
UPO	Uncertain Portfolio Optimization
URAP	University Ranking by Academic Performance
URL	Uniform Resource Locator
US	United States
USA	United States of America
USD	United States Dollar currency
USEDEA	Uncertain Super-Efficiency Data Envelopment Analysis
USSR	Union of Soviet Socialist Republics
VADER	Valence Aware Dictionary and Sentiment Reasoned
VAR	Vector Autoregressive Model
VECM	Vector Error Correction Model
VRS	Variable Returns to Scale
WASPAS	Weighted Aggregated Sum Product Assessment
WGI	Worldwide Governance Indicators
WSI	World Sustainable Indicators
WTO	World Trade Organization.
WWI	The First World War
XGBoost	Extreme Gradient Boosting
ZINB	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson

# The Cobb-Douglas Production Function for an Exponential Model



Roman G. Smirnov, Kunpeng Wang, and Ziwei Wang

**Abstract** We investigate China's post-1978 economic data in terms of compatible Cobb-Douglas production functions exhibiting different properties for different periods of time. Our methodology is grounded in the fact that the Cobb-Douglas function can be derived under the assumption of exponential growth in production, labor, and capital. We show that it appears to be the case by employing R programming and the method of least squares. Each Cobb-Douglas function used to characterize the economic growth within the corresponding period of time is determined by specifying the values of the labor share from the available empirical data for the period in question. We conclude, therefore, that the Cobb-Douglas function can be employed to describe the growth in production for the periods 1978–1984, 1985–1991, 1992–2002, 2003–2009, and 2010–2017 each marked by specific events that impacted the Chinese economy.

**Keywords** China's economic growth · Cobb-Douglas production function · Data analysis · Exponential model · Invariants · Reforms

## 1 Introduction

China's economic growth, spurred by the launch of market-oriented policy reforms in 1978, has been nothing short of spectacular, propelling the country to the position of the world's largest economy (on a purchasing power parity basis). It must be noted,

---

R. G. Smirnov (✉) · Z. Wang  
Department of Mathematics and Statistics, Dalhousie University, 6316 Coburg Road,  
PO Box 15000, B3H 4R2 Halifax, Nova Scotia, Canada  
e-mail: [Roman.Smirnov@dal.ca](mailto:Roman.Smirnov@dal.ca)

Z. Wang  
e-mail: [Ziwei.W@dal.ca](mailto:Ziwei.W@dal.ca)

K. Wang  
Sichuan University-Pittsburgh Institute (SCUPI), Sichuan University,  
610207 Chengdu, Sichuan, China  
e-mail: [kunpeng.wang@scupi.cn](mailto:kunpeng.wang@scupi.cn)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_1](https://doi.org/10.1007/978-3-030-85254-2_1)

however, that although the Chinese economy has been expanding at a steady pace with real annual GDP growth averaging 9.5% through 2017, several notable events that occurred in the last 40 years have arguably impacted its remarkable expansion. First, in October of 1984 the policymakers had introduced the idea of commodity economy (a euphemism of “market economy” at the time) into practice that led to further liberalization of the Chinese economy. Next, in 1992 the reforms had become effectively irreversible, following the implications from Deng Xiaoping’s southern tour. The next important event took place in December of 2001 when China became a member of the World Trade Organization. The effects of the WTO membership had further influenced China’s economic growth. Finally, the year of 2009 had marked China’s remarkable recovery from the Great Recession. Accordingly, we will investigate the economic data from the periods marked by aforementioned events, namely, 1978–1984, 1985–1991, 1992–2001, 2002–2009, 2010–2017.

To conduct our study, we will employ the Cobb-Douglas aggregate production function. Our choice stems from the empirical evidence available to us and analytical properties of this function that connects production and the corresponding impact factors (in most cases, capital and labor). More specifically, we view the Cobb-Douglas functions as a consequence of the vigorous growth in production, labor, and capital. It is normally possible to describe such a growth by exponential models that can be fit to the corresponding data (see Sato (1981), Sato and Ramachandran (2012), Smirnov and Wang (2020a), Smirnov and Wang (2020b) for more details and references). In effect, that is exactly what was done in 1928 by Charles Cobb and Paul Douglas themselves Cobb and Douglas (1928). Importantly, this observation also puts limitations on the applicability of the Cobb-Douglas function in growth models. It is our contention that only the data that can be approximated with exponentials can be accurately described by the Cobb-Douglas function. In what follows, we will demonstrate that China’s economic growth in production, labor, and capital satisfies this requirement, that is the data for each of the five periods outlined above can be approximated by the corresponding exponential functions.

Therefore, our goal in this paper is to use the available empirical data to characterize the economic growth in post-1978 China during each of the five periods from the viewpoint of analytic properties of the Cobb-Douglas function. It should be mentioned that China’s economic data has already been studied with the aid of the Cobb-Douglas function from different perspectives. Thus, for example, Chow and Li Chow and Li (2002) fitted a Cobb-Douglas function to the data from 1950 to 2010 under the assumption of constant return to scales, while Rawski Jefferson et al. (1992) argued that China’s economic data enjoyed decreasing returns to scale during the period 1984–1987. However, in the latter case the corresponding Cobb-Douglas function was defined to depend not only on capital and labor, but also on energy.

## 2 The Method

In this section we briefly discuss the methods used in the paper. We begin by recalling that the Cobb-Douglas production function came into prominence after an economist Paul Douglas and a mathematician Charles Cobb with the aid of statistical analysis came up with an equation describing the relationship among the time series describing the US manufacturing output, labor input, and capital input for the period 1899–1922 Cobb and Douglas (1928). However, it must be mentioned that the function had already gained substantial attention in the years prior to the 1928 paper by Cobb and Douglas (for more details, see Humpfrey Humphrey (1997)). In its most acceptable form, the function is defined by the formula

$$Y = AL^\alpha K^\beta, \quad (1)$$

where  $Y$ ,  $L$ , and  $K$  are production, labor, and capital respectively,  $A > 0$  denotes total productivity, while  $\alpha > 0$  and  $\beta > 0$  are elasticities of labor and capital. In most cases, the derivation of this function has been carried out by various authors either by employing analytical methods (see, for example, Sato Sato (1981)), or mostly through statistical treatment of existing data (see, for example, Cobb and Douglas Cobb and Douglas (1928), Rawski Jefferson et al. (1992), and Chow and Li Chow and Li (2002)). The same observation can be made about the criticisms by the authors who doubt the validity of the Cobb-Douglas function in conjunction with the study of economic data (see, for example, Felipe and Adams Felipe and Adams (2005)). Admittedly, different approaches to the derivation of the Cobb-Douglas function have led to misunderstanding of its true meaning and limitations. In this paper we continue the development of the combined approach proposed by two of the authors (RGS and KW) earlier Smirnov and Wang (2020, a, b) that aims to exploit the natural synergy between the analytical and statistical methods employed in the past to study the Cobb-Douglas function and its properties. Recall that the analytical approach to the development and study of the Cobb-Douglas function is based on the assumption that production, capital, and labor of a given economy grow exponentially, namely, the dynamics is subject to the following simple system:

$$\dot{x}_i = b_i x_i, \quad b_i > 0, \quad i = 1, 2, 3, \quad (2)$$

where  $x_1(t) = L(t)$ ,  $x_2(t) = K(t)$ , and  $x_3(t) = Y(t)$  and the fixed parameters  $b_1$ ,  $b_2$ , and  $b_3$  characterize the corresponding exponential growth in capital, labor, and production as functions of time  $t$ . Then, integrating Eq. (2) to get the solutions

$$x_i = x_i^0 e^{b_i t}, \quad x_i^0, b_i > 0, \quad i = 1, 2, 3 \quad (3)$$

and eliminating  $t$  leads to the derivation of the Cobb-Douglas function as a time-independent invariant of the flow generated by (2) subject to an additional linearity condition:

$$a_1 b_1 + a_2 b_2 + a_3 b_3 = 0, \quad (4)$$

for some parameters  $a_1, a_2, a_3 \in \mathbb{R}$ . Indeed, the product

$$x_1^{a_1} x_2^{a_2} x_3^{a_3} = (x_1^0)^{a_1} (x_2^0)^{a_2} (x_3^0)^{a_3} e^{(a_1 b_1 + a_2 b_2 + a_3 b_3)t}$$

is a time-independent invariant, provided the linearity condition (4) holds. Here  $x_i^0$ ,  $i = 1, 2, 3$  are the initial conditions. Solving the equation  $x_1^{a_1} x_2^{a_2} x_3^{a_3} = C$  for  $x_3$ , we arrive at the Cobb-Douglas function of the form

$$x_3 = C^{\frac{1}{a_3}} x_1^{-\frac{a_1}{a_3}} x_2^{-\frac{a_2}{a_3}}, \quad (5)$$

where the constant  $C = (x_1^0)^{a_1} (x_2^0)^{a_2} (x_3^0)^{a_3}$ . Therefore, the Cobb-Douglas function in this context is consequence of exponential growth in production and the input factors (capital and labor), as well as the condition (4). Moreover, in view of (4) these conditions yield in fact a *family* of Cobb-Douglas functions, namely,

$$x_3 = C^{\frac{1}{a_3}} x_1^{\frac{b_3}{b_1} + \frac{a_2}{a_3} \frac{b_2}{b_1}} x_2^{-\frac{a_2}{a_3}}, \quad (6)$$

or, identifying  $x_1 = L$ ,  $x_2 = K$ ,  $x_3 = Y$ ,  $C^{\frac{1}{a_3}} = A$ , and  $\frac{a_2}{a_3} = \ell$ , we have

$$Y = AL^{\frac{b_3}{b_1} + \frac{b_2}{b_1} \ell} K^{-\ell}. \quad (7)$$

We note that the exponential growth in the variables  $x_1 = L$ ,  $x_2 = K$ , and  $x_3 = Y$ , as well as the linearity condition (4) make perfect sense from the economic standpoint. Indeed, they simply mean that on the one hand the economy is undergoing a robust growth represented by the growth in production, labor, and capital. On the other hand the variables are not independent. Specifically,  $\ln x_1$ ,  $\ln x_2$ , and  $\ln x_3$  are linearly connected via the condition (4) and the formula (6). This is also acceptable, because the three variables describe the same economy in which a change in one of the variables inevitably yields the corresponding change in the other two. One can assert that multicollinearity in this case is not a bag, it is a feature. For example, an increase in capital will affect both production and labor.

In order to specify the parameter  $\ell$  in (7), we need an extra condition. One such a condition is the assumption of constant returns to scale (i.e., the condition  $\alpha + \beta = 1$  in (1) and  $\frac{b_3}{b_1} + \ell \frac{b_2}{b_1} - \ell = 1$  in (7), which yields the following Cobb-Douglas function

$$Y = AL^{\frac{b_3 - b_2}{b_1 - b_2}} K^{\frac{b_3 - b_1}{b_2 - b_1}}. \quad (8)$$

We note that the function (8) is economically sound, provided the elasticities of the inputs are positive, which implies that either  $b_2 > b_3 > b_1$ , or  $b_1 > b_3 > b_2$ . This is a further limitation on the economic growth determined by (2) with the meaning that

the function (8) enjoys constant returns to scale iff production does not grow faster (slower) than both labor and capital.

Another well-known approach to the derivation of the Cobb-Douglas function (1) is statistical in nature and rooted in the study of the available data representing the growth in production, labor, and capital for a given economy. For example, that is exactly how Charles Cobb and Paul Douglas themselves employed the function (1) as the basis of a statistical procedure for estimating the relationship between production, labor, and capital. Specifically, they employed data from the US manufacturing sector for 1899–1922, assuming constant return to scale in (1), to fit the corresponding function  $Y = AL^{1-\beta}K^\beta$  with the aid of statistical analysis to this data. The value for the elasticity of labor was found to be 0.75, while the National Bureau of Economic Research determined this value empirically to be 0.741 Cobb and Douglas (1928). In spite of the fact that this approach was later employed with much success to study other economic data sets (see Douglas Douglas (1976) and the relevant references therein), the question remained: *Can the data studied by Cobb and Douglas be fitted with another function of the type (1) that does not enjoy constant return to scale?* In view of the above observations, the answer to this question is *yes*, which was confirmed in Smirnov and Wang (2020a) by employing a modification of the statistical method used by Cobb and Douglas originally that incorporated the analytical approach briefly outlined above. More specifically, we proposed in Smirnov and Wang (2020a, b) the following approach. Given economic data representing growth in production, labor, and capital, we first verify whether the exponential model (3) can be fitted to the data, using statistical methods such as R programming and the method of least squares. If it is the case, we determine for each variable  $x_i$ ,  $i = 1, 2, 3$  the corresponding initial values  $x_i^0$  and the parameters  $b_i$ ,  $i = 1, 2, 3$  representing exponential growth. Now we know for a fact that fitting a Cobb-Douglas function to the data is possible. Furthermore, if (the approximate values of) the parameters  $b_i$ ,  $i = 1, 2, 3$  satisfy either the inequality  $b_2 > b_3 > b_1$ , or  $b_1 > b_3 > b_2$  the Cobb-Douglas function of the form (8) can be fitted to the given data. Otherwise, the elasticities of capital and labor in (1) satisfy either  $\alpha + \beta < 1$ , or  $\alpha + \beta > 1$  for any element of the family (7). In the former case we have decreasing returns to scale determined by the inequalities  $b_3 < b_2$  and  $b_3 < b_1$ , while in the latter—increasing returns to scale determined by the inequalities  $b_3 > b_2$  and  $b_3 > b_1$  for *all* elements of the family of the Cobb-Douglas functions (7) (see Smirnov and Wang (2020b) for more details).

It must be mentioned that the data originally studied by Cobb and Douglas in Cobb and Douglas (1928) has been further investigated using the algorithm outlined above Smirnov and Wang (2020a). Specifically, we have verified, using R, that indeed the time series representing the changes in production, labor, and capital approximately followed exponential growth with  $b_1 = 0.025496$  (labor),  $b_2 = 0.064725$  (capital), and  $b_3 = 0.035926$  (production). Therefore, the family of the Cobb-Douglas functions compatible with the given data is determined from the formula (7) to be

$$Y = AL^{1.409084+2.538634\ell}K^{-\ell}. \quad (9)$$

Furthermore, we have  $b_2 > b_3 > b_1$  and so constant return to scale is possible, that is the family of the Cobb-Douglas functions (9) contains the element of the form (8). Substituting these values for  $b_1$ ,  $b_2$ , and  $b_3$  into the formula (8), we found the elasticity of labor to be approximately 0.734125, which was very close to the value determined by Cobb and Douglas in Cobb and Douglas (1928) directly *under the assumption of constant return to scale*. However, this is not the only Cobb-Douglas function that affords a good fit to the data. For example, the function  $Y = 0.471016LK^{0.161149}$ , which is an element of the family (9) also affords a good fit to the data studied in Cobb and Douglas (1928). However, it no longer enjoys constant returns to scale. Nevertheless, the Cobb-Douglas function enjoying constant return to scale derived in Cobb and Douglas (1928) is the right choice because its value of the elasticity of labor was independently confirmed by the National Bureau of Economic Research. In what follows, we apply this method to study the Chinese economic data from the period 1978–2017.

### 3 The Data

This section focuses on China's economic data from the period 1978–2017. Our goal here is to collect, unify and tabulate the data representing China's growth in production, labor, and capital for this period. We first gather the data representing the nominal GDP for the period 1978–98 from Table 3-1 in China's Statistical Yearbook (CSY), Volume 2020. Next, we employ Chow and Li's method Chow and Li (2002) to obtain the production series data in 1978 prices through dividing nominal GDP by an adjusted deflator. The deflator for the period 1978–2017 employed here comes from the website Indexmundi (<https://www.indexmundi.com/facts/china/gdp-deflator>); it is given in terms of index values with the value at 1978 taken as 100%, which is consistent with implicit price deflator in Chow and Li (2002). The labor force figures from Table 2-10 in the CSY, Volume 2020 is used as the labor input. The capital time series data for 1978–1998 has been gathered and tabulated by Chow and Li Chow and Li (2002). We compute the capital series data after 1998 based on the capital stock values from 1978 to 1998 used by Chow and Li (see Table 1 in Chow and Li (2002)), employing their formula

$$K_t = K_{t-1} + RNI_t, \quad (10)$$

where  $K$  is capital and  $RNI$  represents real net investment (see Chow and Li (2002) for more details and references). The values of  $RNI$  are calculated using gross investment, net investment (gross investment less total provincial depreciation), and real gross investment (deflated gross investment in 1978 prices). We find gross investment and total provincial depreciation based on items in the tables of the GDP data by expenditure approach at provincial level published in the CSY, Volumes 1999–2012. It must be noted that the total provincial depreciation data for 2004, 2008, and 2013 are not presented in the CSY and so we estimate the data by averaging the total



provincial depreciation values at their consecutive years. For example, depreciation for 2014 is obtained by finding the mean of values for the years 2013 and 2015.

We normalize the time series data by using dimensionless index values with values at 1978 taken as 100. We present the index values of capital, labor, and production from 1972 to 2017 in Table 3 on a logarithmic scale (see Appendix A). Next, we break the data from the period 1978–2017 into the following five data sets: 1978–1984, 1985–1991, 1992–2001, 2002–2009, and 2010–2017. As was already mentioned, each period is marked by specific events that significantly influenced China's economy.

## 4 The Results

Taking the logarithm of both sides in (3), we linearize the variables as follows:

$$\ln x_i = C_i + b_i t, \quad i = 1, 2, 3, \quad (11)$$

where  $C_i = \ln x_i^0$ ,  $x_1 = L$  (labor),  $x_2 = K$  (capital), and  $x_3 = Y$  (production).

Next, we recover the corresponding values of the coefficients  $C_i$ ,  $b_i$ ,  $i = 1, 2, 3$  for Sets 1–5. Employing R and the method of least squares, we arrive at the following values.

Set 1 (1978–1984):

$$\begin{aligned} b_1 &= 0.030814, \quad C_1 = 4.599695 \text{ (labor)}, \\ b_2 &= 0.062199, \quad C_2 = 4.593280 \text{ (capital)}, \\ b_3 &= 0.084770, \quad C_3 = 4.583829 \text{ (production)}. \end{aligned} \quad (12)$$

Set 2 (1985–1991):

$$\begin{aligned} b_1 &= 0.047535, \quad C_1 = 4.794706 \text{ (labor)}, \\ b_2 &= 0.096918, \quad C_2 = 5.078597 \text{ (capital)}, \\ b_3 &= 0.077071, \quad C_3 = 5.283842 \text{ (production)}. \end{aligned} \quad (13)$$

Set 3 (1992–2001):

$$\begin{aligned} b_1 &= 0.010925, \quad C_1 = 5.102944 \text{ (labor)}, \\ b_2 &= 0.108844, \quad C_2 = 5.773019 \text{ (capital)}, \\ b_3 &= 0.090424, \quad C_3 = 5.911680 \text{ (production)}. \end{aligned} \quad (14)$$

Set 4 (2002–2009):

$$\begin{aligned} b_1 &= 0.004855, C_1 = 5.208945 \text{ (labor)}, \\ b_2 &= 0.127195, C_2 = 6.841267 \text{ (capital)}, \\ b_3 &= 0.109907, C_3 = 6.781214 \text{ (production)}. \end{aligned} \quad (15)$$

Set 5 (2010–2017):

$$\begin{aligned} b_1 &= 0.002979, C_1 = 5.246106 \text{ (labor)}, \\ b_2 &= 0.120791, C_2 = 7.907700 \text{ (capital)}, \\ b_3 &= 0.074385, C_3 = 7.644393 \text{ (production)}. \end{aligned} \quad (16)$$

We verify that the errors, represented by the \$values in each estimation, are all less than 1, which suggests that the formulas (11) fit quite well to the data in Table 3 (all of the R programming codes used here are available upon request). We also obtain satisfactory values of the goodness of fit in each regression. Therefore, we arrive at the following families of Cobb-Douglas functions (7) associated with each set.

$$\begin{aligned} \text{Set 1 (1978–1984): } Y_1 &= AK^{2.751022+2.018531\ell} L^{-\ell}, \\ \text{Set 2 (1985–1991): } Y_2 &= AK^{1.621353+2.038877\ell} L^{-\ell}, \\ \text{Set 3 (1992–2001): } Y_3 &= AK^{8.276796+9.962838\ell} L^{-\ell}, \\ \text{Set 4 (2002–2009): } Y_4 &= AK^{22.637899+26.214963\ell} L^{-\ell}, \\ \text{Set 5 (2010–2017): } Y_5 &= AK^{24.969789+40.547499\ell} L^{-\ell}. \end{aligned} \quad (17)$$

We see that only the first family of Cobb-Douglas functions, corresponding to Set 1, does not contain the Cobb-Douglas function satisfying the condition of constant returns to scale, which is due to the inequality  $b_3 > b_2 > b_1$ . Next, in order to determine an appropriate element within each of the families (17), we will use additional empirical characterizations of each set. Recall that Cobb and Douglas in Cobb and Douglas (1928) verified the input elasticities for the function that they derived by comparing the value of  $\alpha$  that they found its empirical value determined by the National Bureau of Economic Research. In what follows, we will employ a similar approach. Specifically, we make use of the fact that the labor elasticity  $\alpha$  in (1) represents the (constant) value of the labor share, that is

$$\alpha = \frac{\partial Y}{\partial L} \frac{L}{Y},$$

which is compatible with the formula for the labor share derived in Smirnov and Wang (2020) bypassing the Cobb-Douglas function. To fix  $\alpha$  and thus  $\beta$  and  $A$  within each set, we use this fact and the available empirical data to compute  $\alpha$  directly, assuming that within each Set 1–5 the value of labor share is constant. It must be mentioned that

**Table 1** China’s labor share data by income approach from 2008–2017

Year	2008	2009	2010	2011	2012
Labor share	None	0.4662	0.4501	0.4494	0.4559
Year	2013	2014	2015	2016	2017
Labor share	None	0.4651	0.4789	0.4746	0.4751

**Table 2** The values of parameters  $\alpha$ ,  $\beta$  and  $A$  that determine the corresponding Cobb-Douglas production functions for each of the five periods

Period	$\alpha$	$\beta$	$A$	Error
1978–1984	0.594357	1.068429	0.046888	0.093819
1985–1991	0.598100	0.501866	0.875754	0.011126
1992–2001	0.579310	0.772623	0.222041	0.001630
2002–2009	0.494541	0.845208	0.206569	0.002338
2010–2017	0.464148	0.604368	1.537645	0.000108

although the recent empirical results Bentolila and Saint-Paul (2003), Bentolila and Saint-Paul (2003) show that the labor share is not constant at least in the medium run, in the short run this is a reasonable assumption. Indeed, the data studied in Chong-En and Zhenjie (2010), Qi (2020) have shown that China’s labor share declined substantially between 1978 and the late 2000s/2010s, but roughly remained constant during the five aforementioned (short) periods of time. The labor share values for 1978–2007 are taken from the combined labor share in Table 1 presented in Chong-En and Zhenjie (2010). We compute China’s labor share between 2008 and 2017 by employing the income approach. Thus, we have computed the provincial compensation of employees from the CSY (Vol.2008-Vol.2018) and obtained the values of annual labor share by dividing aggregate provincial compensation by nominal GDP (see Table 2). Then, we have determined the output elasticity  $\alpha$  by calculating the mean values of labor share in each period. Note the provincial compensation data in 2008 and 2013 are not released, but this does not affect the mean values significantly.

As follows from our definition of the Cobb-Douglas function, constant returns to scale are not an intrinsic property of the Cobb-Douglas function that fits to given data. We substitute the above values of  $\alpha$  into the formulas for each of the five families (17), thus determining the parameter  $\ell$  leading to the the corresponding values of  $\alpha$  and  $\beta$ . Next, we find the values of total factor productivity  $A$  employing the Brent regression method. We summarize the results in Table 2.

## 5 Concluding Remarks

Using the notion of the family of Cobb-Douglas functions given by (7) that is determined by the input and output variables exhibiting exponential growth (2), we were able to describe the growth of China's economy during each of the five consecutive periods from 1978 till 2017. In particular, we see that the growth in production was the strongest vs the growth in capital and labor from 2010 to 2017 (see Set 5 in (17)). During this period the highest was also the total productivity factor  $A$ . We also note that the growth in labor was the slowest during this last period. Overall, the growth in labor during the whole period appears to be logistic rather than exponential, which makes a perfect sense.

In summary, our model is based upon the following algorithm employed in this paper to study China's economy.

First, we check, using R, whether the output and input parameters can be accurately approximated by exponential functions. It is the case, we derive the corresponding parameters  $b_1$ ,  $b_2$ , and  $b_3$  representing exponential growth in each of the variables.

Next, we derive the corresponding family of the Cobb-Douglas functions (7). The problem is not solved yet, because we need some additional information needed to completely fix the parameters  $A$ ,  $\alpha$ , and  $\beta$ . To do this we make use of additional empirical data—in this case the labor share—to derive the Cobb-Douglas function that connects the output and inputs during a given period. The parameters  $A$ ,  $\alpha$ , and  $\beta$  are used to better understand and characterize the growth of production vs capital and labor for a given economy.

Our approach, which incorporate both statistical and mathematical methods, is a generalization of the approach employed by Cobb and Douglas in 1928. Indeed, employing our method, we can not only pick a Cobb-Douglas function that is a good fit for a data set representing economic growth, but we can also pick the right function and explain why there are other Cobb-Douglas functions that are compatible with a given data, which, nonetheless, do not relate the output and input variables for a given economy.

## Appendices

### *A The Time Series Data from 1978–2017*

See Table 3.

**Table 3** The time series data from 1978–2017

Year	Production $Y$	Labor $L$	Capital $C$
1978	4.605170	4.605170	4.605170
1979	4.678628	4.626655	4.658306
1980	4.753533	4.658726	4.714056
1981	4.803614	4.690418	4.765696
1982	4.890866	4.725695	4.829077
1983	4.99653	4.750573	4.902155
1984	5.138630	4.787795	4.984775
1985	5.262161	4.821978	5.076008
1986	5.345325	4.849838	5.172084
1987	5.455029	4.878687	5.269186
1988	5.562040	4.907648	5.376683
1989	5.603906	4.925795	5.477815
1990	5.644398	5.083016	5.563794
1991	5.732475	5.094411	5.649897
1992	5.864016	5.104453	5.745329
1993	5.991187	5.114321	5.872274
1994	6.114023	5.123959	5.996493
1995	6.203785	5.132961	6.117242
1996	6.299963	5.145880	6.230647
1997	6.391264	5.158418	6.334437
1998	6.461915	5.170052	6.432802
1999	6.538906	5.180712	6.531172
2000	6.621475	5.190344	6.630719
2001	6.699336	5.200173	6.737104
2002	6.790904	5.206786	6.849440
2003	6.890539	5.212989	6.971409
2004	6.990493	5.220124	7.096924
2005	7.093971	5.225268	7.216096
2006	7.218866	5.229693	7.337764
2007	7.356048	5.234257	7.468736
2008	7.452116	5.237478	7.604464
2009	7.534165	5.240966	7.746788
2010	7.632617	5.244612	7.889437
2011	7.718726	5.248742	8.025847
2012	7.801369	5.252452	8.156867
2013	7.870421	5.256005	8.282785
2014	7.953807	5.259584	8.405889
2015	8.016168	5.262143	8.519031
2016	8.085520	5.264104	8.627565
2017	8.159216	5.264581	8.736350

## References

- Bentolila S, Saint-Paul G (2003) Explaining movements in the labor share. *BE J Macroecon* 3(1):1–33
- Bergeaud A, Cette G, Lecat R (2017) Total factor productivity in advanced countries: a longterm perspective. *Int Prod Monitor* 32:6–24
- Chong-En B, Zhenjie Q (2010) The factor income distribution in China: 1978–2007. *China Econ Rev* 21(4):650–670
- Chow GC, Li KW (2002) China’s economic growth: 1952–2010. *Econ Devel Cult Change* 51(1):247–256
- Cobb CW, Douglas PH (1928) A theory of production. *Am Econ Rev* 18(Suppl):139–165
- Douglas PH (1976) The Cobb-Douglas production function once again: Its history, its testing, and some new empirical values. *J Polit Econ* 84:903–915
- Felipe J, Adams FG (2005) “A theory of production” the estimation of the Cobb-Douglas function: a retrospective view. *Eastern Econ J* 31(3):427–445
- Humphrey TM (1997) Algebraic production functions and their uses before Cobb-Douglas. *FRB Richm Econ Quart* 83(1):51–83
- Jefferson GH, Rawski TG, Zheng Y (1992) Growth, efficiency, and convergence in China’s state and collective industry. *Econ Devel Cult Change* 40(2):239–266
- Mendershausen H (1938) On the significance of Professor Douglas’ production function. *Econometrica* 6(2):143–156
- Qi H (2020) Power relations and the labour share of income in China. *Cambridge J Econ* 44(3):607–628
- Raurich X, Sala H, Sorolla V (2012) Factor shares, the price markup, and the elasticity of substitution between capital and labor. *J Macroecon* 34(1):181–198
- Sato R (1981) *Theory of technical change and economic invariance*. Academic Press, New York
- Sato R, Ramachandran RV (2012) *Symmetry and economic invariance: an introduction*. Springer Science & Business Media, New York
- Smirnov RG, Wang K (2020) On the validity of the concept of a production function in economics. Preprint
- Smirnov RG, Wang K (2020) The Cobb-Douglas function revisited. Accepted for publication in the Proceedings of The V AMMCS International Conference. <https://arxiv.org/abs/1910.06739>
- Smirnov RG, Wang K (2020) In search of a new economic model determined by logistic growth. *European J Appl Math* 31(2):339–368

# Threshold Unit Root Tests with Smooth Transitions



Mehmet Özcan  and Funda Yurdakul 

**Abstract** Since threshold autoregressive models were discovered, many unit root tests have been developed to test the unit root null hypothesis when considering regime change. On the other hand, Sollis, *J Time Ser Anal* 25:409–417, 2004, indicates that a threshold unit root test could be combined with some smooth transition logistic functions which are introduced by Leybourne et al., *J Time Ser Anal* 19:83–97, 1998. This paper investigates whether the Caner and Hansen, *Econometrica* 69:1555–1596, 2001, unit root test could be expanded with smooth transition functions and demonstrates the performance of this new unit root testing process with Monte Carlo simulations. Simulation results for finite sample properties show reasonable empirical size and power values. Also, the proposed unit root testing procedure is used to test unit root null hypothesis for industrial production indices of United States of America and Turkey.

**Keywords** Threshold model · Smooth transition · Unit root · Monte Carlo simulations

## 1 Introduction

The threshold autoregressive (TAR) models which were introduced by Tong (1978) make it possible to estimate regime switching dynamics of time series. These models also provide new developments for unit root testing studies. Enders and Granger (1998) was the first paper that propose new threshold unit root test statistics which considers regime switching in time series. In addition to this, Caner and Hansen

---

M. Özcan (✉)

Department of Economics, Faculty of Economics and Administrative Sciences, Karamanoglu Mehmetbey University, Karaman, Turkey  
e-mail: [mehmetozcan@kmu.edu.tr](mailto:mehmetozcan@kmu.edu.tr)

F. Yurdakul

Department of Econometrics, Faculty of Economics and Administrative Sciences, Ankara Hacı Bayram Veli University, Ankara, Turkey  
e-mail: [funda.yurdakul@hbv.edu.tr](mailto:funda.yurdakul@hbv.edu.tr)

(2001) employed another threshold unit root testing procedure which has more detailed alternative hypotheses and more comprehensive threshold autoregressive models than Enders and Granger (1998). Moreover, Leybourne et al. (1998) (Henceforth LNV) offered to use logistic smooth transition functions for modeling structural changes and expand Augmented Dickey Fuller (ADF) (Dickey and Fuller 1981) unit root testing procedure with nonlinear smooth transition functions. This paper proposes a new unit root testing procedure to test unit root null against the alternative hypothesis for asymmetric adjustment around a smooth transition between two linear trends. To achieve this, a combination strategy that was developed by Sollis (2004) is adopted. Sollis (2004) offered to combine smooth transition methodology employed by LNV with the Enders and Granger (1998) threshold unit root test that allows stationary asymmetric adjustment around a smooth transition between linear trends under the alternative hypothesis. In the same way, this study keeps LNV approach with a different estimation method for dealing with smooth transition. However, the Enders and Granger (1998) threshold unit root test was changed with Caner and Hansen (2001). In order to be sure about performance of the new combined tests, the size and power of the test are investigated with Monte Carlo simulation experiments. The results of these experiments offer reasonable size and power values. Moreover, industrial production index series of the United States of America (USA) and Turkey are considered for empirical application. Conforming to application results, it can be said that these two series carry different characteristics. Remarkable structural change in 2008 could be observed in estimated smooth transition series of both industrial production indices. Comparing the two series, the residual series of smooth transition models for Turkey does not have statistically significant threshold effect (regime switching behavior), while the residual series of USA does have a statistically significant threshold effect. Therefore, these two series are good choices for empirical application of combined tests.

The following is an outline of the paper. Section 2 contains detailed explanations about LNV, Caner and Hansen (2001), and combined tests. Simulation studies, which are used to investigate finite sample properties of the combined tests, are reported in Sect. 3. Empirical application is placed in Sect. 4. Section 5 gives some concluding remarks.

## 2 Smooth Transition Models, Unit Root Test, and Null Hypothesis

Caner and Hansen (2001) developed a unit root test that allowed the unit root null to be tested against some alternatives of stationary with two regime threshold autoregressive processes. General form of the threshold model, which is considered for unit root testing, could be expressed as follows:

$$\Delta y_t = \theta'_1 x_{t-1} 1_{\{z_{t-1} < \lambda\}} + \theta'_2 x_{t-1} 1_{\{z_{t-1} \geq \lambda\}} + e_t, \quad (1)$$



$t = 1, \dots, T$ , where  $x_{t-1} = (y_{t-1}r' \Delta y_{t-1} \dots \Delta y_{t-k})'$ ,  $z_t = y_t - y_{t-m}$  is a threshold variable for delay parameter  $m \geq 1$ ,  $r_t$  is a vector of deterministic components,  $1_{\{\cdot\}}$  is the indicator function, and  $e_t$  is an independent and identically distributed (iid) error. It should be noted that  $\theta_1$  and  $\theta_2$  contain coefficients of the first and second regime, respectively, as follows:

$$\theta_1 = \begin{pmatrix} \rho_1 \\ \beta_1 \\ \alpha_1 \end{pmatrix}, \quad \theta_2 = \begin{pmatrix} \rho_2 \\ \beta_2 \\ \alpha_2 \end{pmatrix},$$

where  $\rho_1$  and  $\rho_2$  are the slope coefficients of  $y_{t-1}$ , and  $\alpha_1$  and  $\alpha_2$  are the  $k$  dimension vectors and are the slope coefficients of  $\Delta y_{t-1}, \dots, \Delta y_{t-k}$ . Lastly,  $\beta_1$  and  $\beta_2$  have the same dimension as  $r_t$ , and they represent coefficients of deterministic components. Moreover, threshold value  $\lambda$  is unknown, and it should be estimated. It is a value of the threshold variable. According to estimation strategy, and the values of the threshold variable are sorted from smallest to largest. Let us call this series as  $\Lambda$ . Then, some observations of a certain proportion are discarded from the beginning and end of  $\Lambda$ . This specific proportion is called the trimming rate, and it is represented by  $\pi$ . Each of the remaining values ( $\Lambda_i$ ,  $i = (\pi 100), \dots, T(1 - \pi)$ ) is considered as a potential candidate for the threshold value, and Model (1) could be estimated with each of these values by ordinary least squares (OLS) as follows:

$$\Delta y_t = \hat{\theta}_1(\Lambda_i) x_{t-1} 1_{\{z_{t-1} < \Lambda_i\}} + \hat{\theta}_2(\Lambda_i) x_{t-1} 1_{\{z_{t-1} \geq \Lambda_i\}} + \hat{e}_t(\Lambda_i). \quad (2)$$

Estimation of  $\hat{\sigma}^2$  for each  $\Lambda_i$  could be written from estimation of (2) as follows:

$$\hat{\sigma}^2(\Lambda_i) = \sum_{t=1}^T \hat{e}_t(\Lambda_i)^2 / T, \quad (3)$$

$t = 1, \dots, T$ . Threshold value  $\lambda$  estimation is defined as follows:

$$\hat{\lambda} = \arg \min \hat{\sigma}^2(\Lambda_i). \quad (4)$$

Unlike to Enders and Granger (1998), Caner and Hansen (2001) allowed for estimation of threshold parameter and expansion of the model with deterministic components and lagged dependent variables. Therefore, Caner and Hansen (2001) offered a more comprehensive model and estimation strategy.

Petrucelli and Woolford (1984) indicated that  $y_t$  is stationary in a first order threshold autoregressive model if  $\rho_1 < 0$ ,  $\rho_2 < 0$ , and  $(1 + \rho_1)(1 + \rho_2) < 1$ . For this reason, the null hypothesis of unit root,  $H_0 : \rho_1 = \rho_2 = 0$ , should be tested against  $H_1 : \rho_1 < 0$  and  $\rho_2 < 0$ . However, it should be noted that the F test, which is proposed by Enders and Granger (1998), rejects null  $H_0$  against  $H_1 : \rho_1 \neq 0$  and  $\rho_2 \neq 0$ . Consequently, Caner and Hansen (2001) proposed four alternative hypotheses and two test statistics in order to get a more advance testing process. These alternatives could be written as follows:

$$H_{10} : \rho_1 \neq 0 \text{ and/or } \rho_2 \neq 0, \quad (5)$$

$$H_{20} : \rho_1 < 0 \text{ and } \rho_2 < 0, \quad (6)$$

$$H_{21} : \rho_1 < 0 \text{ and } \rho_2 = 0, \quad (7)$$

$$H_{22} : \rho_1 = 0 \text{ and } \rho_2 < 0. \quad (8)$$

Equation (5) is defined as an unrestricted stationary alternative. Equations (6), (7), and (8) state restricted stationary alternatives. Further,  $H_{21}$  and  $H_{22}$  indicate a new concept for unit root testing. These alternatives are called partial unit root alternatives. There is no doubt that, if nonstationary null is rejected against  $H_{21}$  and  $H_{22}$ , there is no way to say  $y_t$  is stationary (Caner and Hansen 2001: 1568). Two different Wald statistics are offered to test the null against these four alternatives. The first one is for unrestricted alternative (5), and it is the two-sided Wald test from (1) as follows:

$$R_{2T} = t_1^2 + t_2^2. \quad (9)$$

According to the OLS estimation from (2),  $t_1$  and  $t_2$  are the t ratios of  $\hat{\rho}_1$  and  $\hat{\rho}_2$ , respectively. Conversely, the second test is a one-sided Wald test, and it has power against the restricted alternatives (6), (7), and (8) as follows:

$$R_{1T} = t_1^2 1_{\{\hat{\rho}_1 < 0\}} + t_2^2 1_{\{\hat{\rho}_2 < 0\}}. \quad (10)$$

It should be emphasized that  $R_{1T}$  is unable to distinguish among the alternatives (6), (7), and (8). Caner and Hansen (2001) offers a solution, which is based on the significance of individual t statistics. If both test statistics  $-t_1$  and  $-t_2$  are statistically significant, the null is rejected against (6). On the contrary, if only one of  $-t_1$  or  $-t_2$  is statistically significant, this situation points to a partial unit root case, which is represented by (7) or (8). Another essential point is the distributions of test statistics under null. Because of the null hypothesis ( $H_0 : \rho_1 = \rho_2 = 0$ ), distributions are required to be investigated under two cases. If there is a threshold effect ( $\theta_1 \neq \theta_2$ ),  $\lambda$  is identified and this case is called identified threshold. On the other hand, if coefficients of two regime are equal to each other ( $\theta_1 = \theta_2$ ),  $\lambda$  is not identified, and this is a case of an unidentified threshold. To address this, Caner and Hansen investigated asymptotic and bootstrap distributions of  $R_{1T}$  and  $R_{2T}$ . According to simulation results, it was found that bootstrap methods are superior to asymptotic approximations. Moreover, bootstrap distributions of  $R_{1T}$  and  $R_{2T}$  show different characteristics under identified and unidentified threshold cases. Caner and Hansen suggested using unidentified threshold bootstrap method to find p-values.

As mentioned in Sollis (2004), the LNV test is a nonlinear alternative to the structural break tests of Perron (1989) and Zivot and Andrews (1992), which allows

for stationary autoregressive processes with smooth transitions under alternatives. Instead of using dummy variables, the LNV approach benefits smooth transition models to fit both instantaneous and gradual breaks. Three smooth transition models are introduced for a time series  $y_t$  by LNV:

$$\text{Model A} : y_t = \delta_1 + \delta_2 S_t(\gamma, \tau) + v_t, \quad (11)$$

$$\text{Model B} : y_t = \delta_1 + \varphi_1 t + \delta_2 S_t(\gamma, \tau) + v_t, \quad (12)$$

$$\text{Model C} : y_t = \delta_1 + \varphi_1 t + \delta_2 S_t(\gamma, \tau) + \varphi_2 t S_t(\gamma, \tau) + v_t, \quad (13)$$

where  $v_t$  is an I(0) error with zero mean, and  $S_t(\gamma, \tau)$  is the logistic function, which could be written as follows:

$$S_t(\gamma, \tau) = (1 + \exp\{-\gamma[t - \tau T]\})^{-1}, \quad (14)$$

where  $\gamma$  determines speed of transition and should be greater than zero ( $\gamma > 0$ ).  $\tau$  points the mid-point of the transition. Lastly,  $T$  is the sample size. Two sets of hypotheses could be tested by (11), (12), and (13) with the function of (14) as follows:

$$\begin{aligned} H_0 : y_t &= \mu_t, \quad \mu_t = \mu_{t-1} + \varepsilon_t, \\ H_1 : \text{Stationary } y_t &\text{ with (11), (12), (13),} \end{aligned} \quad (15)$$

$$\begin{aligned} H_0 : y_t &= \mu_t, \quad \mu_t = \kappa + \mu_{t-1} + \varepsilon_t, \\ H_1 : \text{Stationary } y_t &\text{ with (12), (13),} \end{aligned} \quad (16)$$

where  $\varepsilon_t$  is the zero mean I(0) error. LNV calculated proper test statistics to test (15) and (16) in a two-step procedure. Firstly, the smooth transition models (11), (12), and (13) are estimated with the nonlinear least square (NLS) estimation method. By doing this, the NLS residuals can be estimated. LNV uses the Broyden, Fletcher, Goldfarb and Shanno (BFGS) algorithm to get NLS residuals from the smooth transition models.

$$\hat{v}_t = y_t - \left( \hat{\delta}_1 + \hat{\delta}_2 S_t(\hat{\gamma}, \hat{\tau}) \right), \quad (17)$$

$$\hat{v}_t = y_t - \left( \hat{\delta}_1 + \hat{\varphi}_1 t + \hat{\delta}_2 S_t(\hat{\gamma}, \hat{\tau}) \right), \quad (18)$$

$$\hat{v}_t = y_t - \left( \hat{\delta}_1 + \hat{\varphi}_1 t + \hat{\delta}_2 S_t(\hat{\gamma}, \hat{\tau}) + \hat{\varphi}_2 t S_t(\hat{\gamma}, \hat{\tau}) \right). \quad (19)$$

Secondly, ADF test statistics are estimated for residuals ( $\hat{v}_t$ ) from (17), (18), and (19) with a Dickey–Fuller type autoregressive model as follows:

$$\Delta \hat{v}_t = \rho \hat{v}_{t-1} + \sum_{i=1}^k \delta_i \Delta \hat{v}_{t-i} + \varphi_t, \quad (20)$$

where  $\varphi_t$  is the stationary error with zero mean for an optimal lag length of  $k$ . LNV offers three test statistics  $s_\alpha$ ,  $s_{\alpha(\beta)}$ , and  $s_{\alpha\beta}$  which are connected with models (11), (12), and (13), respectively. Suitable critical values and finite sample sizes are reported in Leybourne et al. (1998).

The Enders and Granger (1998) and LNV approaches remove deterministic components from the data before testing unit root. By emphasizing this similarity between the two approaches, Sollis (2004) changed the second stage of the test process and suggested the TAR model of Enders and Granger (1998) instead of the standard ADF model as follow:

$$\Delta v_t = I_t \rho_1 v_{t-1} + (1 - I_t) \rho_2 v_{t-1} + \sum_{i=1}^k \xi_i \Delta v_{t-i} + \varpi_t, \quad (21)$$

where  $I_t$  is an indicator function that equals 1 if  $v_{t-1} \geq 0$  or equals 0 if  $v_{t-1} < 0$ . Also,  $\varpi_t$  is a zero mean stationary process. According to Sollis (2004), the error term of (11), (12), and (13) follows the data generation process of (21). Additionally,  $y_t$  is generated by (11), (12), and (13). Therefore,  $y_t$  is called a smooth transition TAR process. The null hypothesis of unit root for  $y_t$  could be defined as  $H_0 : \rho_1 = \rho_2 = 0$ . This null hypothesis is valid regardless of which model  $y_t$  is generated from. There are two different alternative hypotheses against the null of unit root. The first is shown as  $\rho_1 = \rho_2 < 0$  and indicates that  $y_t$  is a symmetric stationary smooth transition TAR process. The second alternative is  $\rho_1 < 0$ ,  $\rho_2 < 0$ , and  $\rho_1 \neq \rho_2$ . This alternative states that  $y_t$  is a stationary smooth transition TAR process with asymmetric adjustment. Sollis (2004) proposes  $F$  statistics to test whether  $y_t$  contains a unit root for  $\rho_1 = \rho_2 = 0$  and  $t$  statistics for  $\rho_1 = 0$  and  $\rho_2 = 0$ . Similar to LNV, Sollis (2004) has three different statistics associated with models (11), (12), and (13):  $(F_\alpha, t_\alpha)$ ,  $(F_{\alpha(\beta)}, t_{\alpha(\beta)})$ , and  $(F_{\alpha\beta}, t_{\alpha\beta})$ , respectively. Finite sample powers and critical values are reported in Sollis (2004).

The key point of the LNV approach is removing the deterministic component of a time series with smooth transition models. After that, the Dickey and Fuller (1979) and/or Enders and Granger (1998) autoregressive models could be estimated for residuals of transition models without constant and trend terms. Similarly, Model (1), which is offered by Caner and Hansen (2001), could be used to test unit root in the second stage of the LNV approach by setting the deterministic component vector equal to zero ( $r_t = \{0\}$ ).

$$\Delta v_t = \Gamma_t \left( \rho_1 v_{t-1} + \sum_{i=1}^k \phi_{1i} \Delta v_{t-i} \right) + (1 - \Gamma_t) \left( \rho_2 v_{t-1} + \sum_{i=1}^k \phi_{2i} \Delta v_{t-i} \right) + \omega_t, \quad (22)$$

where  $\Gamma_t$  is another indicator function that equals 1 if  $\Delta v_{t-d} < \tau$  and equals 0 if  $\Delta v_{t-d} \geq \tau$ .  $\Delta v_{t-d}$  is the threshold variable for (22), with delay parameter  $d$  and  $\tau$  as a threshold value. Both of them should be estimated by the process proposed in Caner and Hansen (2001). Lastly,  $\omega_t$  is a zero mean stationary process. Just as in the Sollis

(2004) and LNV approach, three test statistics could be proposed as  $({}^\alpha R_{1T}, {}^\alpha R_{2T})$ ,  $({}^{\alpha(\beta)} R_{1T}, {}^{\alpha(\beta)} R_{2T})$ , and  $({}^{\alpha\beta} R_{1T}, {}^{\alpha\beta} R_{2T})$ , which are related to (11), (12), and (13), respectively, and all of them are calculated as (9) and (10). Also, the two sets of hypotheses shown in (15) and (16) are identical for the new unit root testing process. However, alternative hypotheses of null  $H_0 : \rho_1 = \rho_2 = 0$  for (22) are defined as (5) for  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{2T}$  and (6), (7), and (8) for  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{1T}$ . As mentioned in Sollis (2004) and Leybourne et al. (1998), if  $v_t$  is stationary, it is acceptable that the  $y_t$  process is also stationary. Furthermore, estimation of smooth transition models (11), (12), and (13) is an important discussion issue. Leybourne et al. (1998) and Sollis (2004) employ BFGS algorithm to solve the NLS estimation problem. However, Vougas (2006) claims that using the sequential quadratic programming (SQP) method for NLS estimation improves the LNV approach. For this reason, the estimation of NLS is done by using the SQP optimization method during this study.

### 3 Finite Sample Properties of the Tests

In order to compute finite sample performance indicators of the new combined unit root tests  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{2T}$ , critical values of the test statistics were calculated by 10,000 Monte Carlo replications under a null model that employed in Sollis (2004) and Leybourne et al. (1998). Table 1 includes the simulated critical values of the null distribution at 0.1, 0.05, and 0.01 significance levels for various sample sizes (T).

Results on Table 1 state that critical value converge to limits with increasing T. Furthermore, compared to the Caner and Hansen (2001)'s critical values, large values

**Table 1** Null critical values of  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta} R_{2T}$  tests at 10, 5, and 1% significance levels

T	${}^\alpha R_{1T}$			${}^{\alpha(\beta)} R_{1T}$			${}^{\alpha\beta} R_{1T}$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
50	19.005	22.448	29.498	26.805	30.884	39.384	30.062	34.411	44.047
100	18.452	21.159	27.696	24.807	27.697	35.419	27.366	30.931	38.650
200	18.278	20.829	26.776	23.491	26.679	33.059	26.431	29.650	36.504
500	17.926	20.515	25.685	22.651	25.434	31.299	25.317	28.117	34.255
T	${}^\alpha R_{2T}$			${}^{\alpha(\beta)} R_{2T}$			${}^{\alpha\beta} R_{2T}$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
50	19.125	22.547	29.719	26.921	30.950	39.504	30.094	34.418	44.047
100	18.562	21.274	27.818	24.886	27.773	35.419	27.407	30.957	38.650
200	18.472	21.071	26.926	23.629	26.800	33.076	26.482	29.693	36.504
500	18.108	20.652	25.898	22.701	25.645	31.354	25.427	28.225	34.255

of  $\alpha, \alpha(\beta), \alpha\beta R_{1T}$  and  $\alpha, \alpha(\beta), \alpha\beta R_{2T}$  statistics are needed to reject null for the new test procedure.

To confirm that the critical values on Table 1 are robust, empirical size values are calculated around 5% nominal size. Size simulation is run by data generation process (DGP) under the null, which is employed by Leybourne et al. (1998) as follows:

$$\Delta y_t = a \Delta y_t + v_t, y_0 = 0, v_t \sim N(0, 1). \quad (23)$$

As seen in (23),  $y_t$  is a random process if  $a = 0$ . Also  $a$  is varied among  $\{-0.8, -0.4, 0, 0.4, 0.8\}$  for size experiment. Rejection frequencies are calculated from 10,000 Monte-Carlo replications. Moreover, the autoregressive lag ( $k$ ) and the delay parameter ( $d$ ) are set to equal 1 for (22). The simulated size values for  $T = \{100, 200\}$  are shown in Table 2.

According to Table 2, size values are significantly close to the nominal size 5%. It is difficult to observe notable size distortions. With this in mind, when  $a > 0$ , increasing the value of  $a$  seems to create over-rejection behavior. However, reduction of the size values is not remarkable enough to worry over the condition of  $a > 0$ .

Power values are another indicator in understanding finite sample performance of test statistics. Unlike the size simulations, power investigation needs a new DGP that produces stationary process. In light of this, a power DGP could be defined as follows:

$$\begin{aligned} y_t &= \delta_1 + \delta_2 S_t(\gamma, \tau) + v_t \quad (\text{Model A}), \\ y_t &= \delta_1 + \varphi_1 t + \delta_2 S_t(\gamma, \tau) + v_t \quad (\text{Model B}), \\ y_t &= \delta_1 + \varphi_1 t + \delta_2 S_t(\gamma, \tau) + \varphi_2 t S_t(\gamma, \tau) + v_t \quad (\text{Model B}), \end{aligned} \quad (24)$$

And

$$\Delta v_t = I_t \rho_1 v_{t-1} + (1 - I_t) \rho_2 v_{t-2} + \epsilon_t \quad \epsilon_t \sim N(0, 1). \quad (25)$$

It is clear that equation group (24) is exactly the same as (11), (12), and (13). Equation (25) is a TAR (1) process with a zero mean stationary  $\epsilon_t$ .  $I_t$  is an indicator function that equals 1 if  $\Delta v_{t-1} < 0$  and equals 0 if  $\Delta v_{t-1} \geq 0$ . Various values are chosen for parameters in (24) for the power simulation experiment.  $\delta_1$  is set at 1.  $\varphi_1$  and  $\varphi_2$  are equal to 0.5. Lastly, the mid-point of transition parameter  $\tau$  is set at 0.5. These are fixed parameters. On the other hand, values of some parameters vary among a set of numbers. For instance,  $\delta_2$  is varied among  $\{2, 10\}$  and the speed of transition parameter  $\gamma$  is taken from values among  $\{0.5, 5\}$ . As indicated in Leybourne et al. (1998) and Sollis (2004), in order to generate stationary  $y_t$ ,  $v_t$  should be a stationary process. Thus, Sollis (2004) employs different combinations of a set of parameter values  $\{-0.1, -0.3, -0.9\}$  for  $\rho_1$  and  $\rho_2$ . However, Caner and Hansen (2001) test statistics also consider a partial stationary process, which is defined with the null hypotheses of (7) and (8). Therefore, it is important to investigate the power of the  $\alpha, \alpha(\beta), \alpha\beta R_{1T}$  and  $\alpha, \alpha(\beta), \alpha\beta R_{2T}$  test statistics under a partial stationary condition.

**Table 2** Empirical sizes of  $\alpha, \alpha^{(\beta)}, \alpha\beta R_{1T}$  and  $\alpha, \alpha^{(\beta)}, \alpha\beta R_{2T}$  unit root tests

		$T = 200$																	
		$T = 100$				Model A				Model B				Model C					
		Model A		Model B		Model C		Model A		Model B		Model C		Model A		Model B		Model C	
$a$		$\alpha R_{1T}$	$\alpha R_{2T}$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$\alpha\beta R_{1T}$	$\alpha\beta R_{2T}$	$\alpha R_{1T}$	$\alpha R_{2T}$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$\alpha\beta R_{1T}$	$\alpha\beta R_{2T}$	$\alpha R_{1T}$	$\alpha R_{2T}$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$\alpha\beta R_{1T}$	$\alpha\beta R_{2T}$
-0.8		5.35	5.48	5.88	6.01	5.10	5.15	5.46	5.31	4.83	4.83	4.78	4.83	5.46	5.31	4.83	4.83	4.78	4.73
-0.4		5.05	5.17	5.52	5.61	5.07	5.10	5.03	4.87	4.76	4.69	4.66	4.69	5.03	4.87	4.76	4.69	4.66	4.69
0		4.78	4.91	5.57	5.65	5.25	5.30	4.91	5.01	4.56	4.53	4.68	4.73	4.91	5.01	4.56	4.53	4.68	4.73
0.4		4.94	4.91	5.43	5.50	4.98	4.98	4.99	4.85	4.58	4.44	4.20	4.23	4.99	4.85	4.58	4.44	4.20	4.23
0.8		4.30	4.25	4.92	4.90	4.39	4.38	4.38	4.25	4.93	4.82	3.77	3.80	4.38	4.25	4.93	4.82	3.77	3.80

Note: Nominal size is 5%

This is why 3 cases are defined to simulate different stationary conditions. The first case considers linear stationary  $v_t$  with a set of parameter values such as  $\rho_1 = \rho_2 = \{-0.1, -0.3, -0.9\}$ . This case simulates the LNV approach. The second case is about the concept of partial stationary. To generate partial stationary  $v_t$ , the autoregressive parameter of one regime has to be equal to 0. Thus, parameter value sets of  $\rho_1$  and  $\rho_2$  are set at  $\rho_1 = \{0\}$  and  $\rho_2 = \{-0.1, -0.3, -0.9\}$ . The third case regards asymmetric stationary  $v_t$  with respect to  $\rho_1 = \{-0.1\}$  and  $\rho_2 = \{-0.1, -0.3, -0.9\}$ . The chosen parameter values under these cases could cover almost all kinds of scenarios that could be faced during applied studies. Power values are calculated from 10,000 Monte Carlo replications. In addition to this, the autoregressive lag ( $k$ ) and the delay parameter ( $d$ ) are set to equal 1 for estimation of the TAR model described in (22). The simulated power values of  $\alpha, \alpha^{(\beta), \alpha\beta} R_{1T}$ ,  $\alpha, \alpha^{(\beta), \alpha\beta} R_{2T}$ , and  $s_{\alpha, \alpha^{(\beta), \alpha\beta}}$  test statistics under the first, second, and third cases are shown in Tables 3, 4, and 5, respectively.

When  $\rho_1 = \rho_2 = \{-0.1, -0.3, -0.9\}$  there is no remarkable difference between power values of the test statistics. On the other hand, the Leybourne et al. (1998) test statistics  $s_{\alpha^{(\beta)}}$  and  $s_{\alpha\beta}$  have slightly more power than  $\alpha^{(\beta), \alpha\beta} R_{1T}$  and  $\alpha^{(\beta), \alpha\beta} R_{2T}$ . As expected, this case represents a linear stationary error term, and the  $\alpha^{(\beta), \alpha\beta} R_{1T}$ ,  $\alpha^{(\beta), \alpha\beta} R_{2T}$ , and  $s_{\alpha, \alpha^{(\beta), \alpha\beta}}$  test statistics show similar performance.

**Table 3** Empirical powers of  $\alpha, \alpha^{(\beta), \alpha\beta} R_{1T}$  and  $\alpha, \alpha^{(\beta), \alpha\beta} R_{2T}$  unit root tests under the case 1  $\rho_1 = \rho_2 = \{-0.1, -0.3, -0.9\}$

$(\delta_2, \gamma)$	$\rho_2$	Model A			Model B			Model C		
		$\alpha R_{1T}$	$\alpha R_{2T}$	$s_\alpha$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$s_{\alpha^{(\beta)}}$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$s_{\alpha\beta}$
(2, 0.5)	-0.1	12.92	12.91	9.05	8.81	8.79	8.87	5.47	5.48	6.78
	-0.3	62.00	61.45	60.14	40.28	40.10	46.36	32.86	32.81	40.79
	-0.9	100.00	100.00	100.00	99.82	99.82	99.93	99.63	99.63	99.91
(2, 5)	-0.1	13.26	13.07	9.21	9.11	9.14	9.11	2.30	2.29	2.37
	-0.3	61.29	60.62	59.10	39.68	39.49	45.38	20.76	20.74	26.33
	-0.9	100.00	100.00	100.00	99.78	99.77	99.94	99.23	99.23	99.81
(10, 0.5)	-0.1	12.62	12.60	8.87	5.90	5.91	5.77	5.45	5.46	6.69
	-0.3	64.54	63.84	62.52	37.37	37.18	42.29	32.95	32.88	40.75
	-0.9	100.00	100.00	100.00	99.85	99.83	99.98	99.63	99.63	99.90
(10, 5)	-0.1	13.12	13.05	8.64	7.34	7.29	5.61	1.99	2.00	2.21
	-0.3	55.22	54.75	52.51	28.40	28.22	32.83	20.59	20.56	26.01
	-0.9	100.00	100.00	100.00	99.76	99.75	99.95	99.24	99.24	99.82

Note Nominal size is 5%, and sample size  $T = 100$



**Table 4** Empirical powers of  $\alpha, \alpha^{(\beta)}, \alpha^\beta R_{1T}$  and  $\alpha, \alpha^{(\beta)}, \alpha^\beta R_{2T}$  unit root tests under the case 2  $\rho_1 = \{0\}$

$(\delta_2, \gamma)$	$\rho_2$	Model A			Model B			Model C		
		$\alpha R_{1T}$	$\alpha R_{2T}$	$s_\alpha$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$s_{\alpha^{(\beta)}}$	$\alpha^\beta R_{1T}$	$\alpha^\beta R_{2T}$	$s_{\alpha^\beta}$
(2, 0.5)	-0.1	9.64	9.85	4.93	7.24	7.22	5.93	3.94	3.99	3.89
	-0.3	50.79	51.09	16.98	28.64	28.73	12.87	21.20	21.44	10.47
	-0.9	99.91	99.91	88.55	99.65	99.63	79.25	99.50	99.50	75.37
(2, 5)	-0.1	9.70	9.89	5.12	7.47	7.46	6.16	1.86	1.88	1.38
	-0.3	50.49	50.83	17.22	28.30	28.50	12.85	13.70	14.11	4.70
	-0.9	99.87	99.85	88.55	99.53	99.54	78.75	98.92	98.90	65.58
(10, 0.5)	-0.1	8.63	8.81	4.41	4.14	4.18	2.98	3.92	3.96	3.74
	-0.3	50.99	51.55	17.09	24.95	25.24	9.74	21.30	21.62	10.45
	-0.9	99.81	99.81	90.07	99.39	99.38	77.35	99.48	99.45	75.45
(10, 5)	-0.1	13.29	13.27	5.87	9.38	9.44	4.77	1.59	1.61	1.09
	-0.3	48.68	49.14	14.33	24.24	24.41	8.74	13.35	13.76	4.34
	-0.9	99.88	99.88	87.29	99.56	99.55	72.20	98.87	98.86	65.25

Note Nominal size is 5%, and sample size  $T = 100$

**Table 5** Empirical powers of  $\alpha, \alpha^{(\beta)}, \alpha^\beta R_{1T}$  and  $\alpha, \alpha^{(\beta)}, \alpha^\beta R_{2T}$  unit root tests under the case 3  $\rho_1 = \{-0.1\}$

$(\delta_2, \gamma)$	$\rho_2$	Model A			Model B			Model C		
		$\alpha R_{1T}$	$\alpha R_{2T}$	$s_\alpha$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$s_{\alpha^{(\beta)}}$	$\alpha^\beta R_{1T}$	$\alpha^\beta R_{2T}$	$s_{\alpha^\beta}$
(2, 0.5)	-0.1	12.92	12.91	9.05	8.81	8.79	8.87	5.47	5.48	6.78
	-0.3	44.66	44.36	29.15	26.33	26.22	21.49	19.76	19.79	17.99
	-0.9	99.88	99.88	96.73	99.37	99.35	91.39	99.20	99.19	88.59
(2, 5)	-0.1	13.26	13.07	9.21	9.11	9.14	9.11	2.30	2.29	2.37
	-0.3	44.39	44.13	29.30	26.46	26.43	21.40	11.32	11.45	9.07
	-0.9	99.90	99.90	96.66	99.28	99.24	90.93	98.75	98.76	81.61
(10, 0.5)	-0.1	12.62	12.60	8.87	5.90	5.91	5.77	5.45	5.46	6.69
	-0.3	45.89	45.65	29.95	22.70	22.63	17.51	19.71	19.74	17.92
	-0.9	99.94	99.94	97.32	99.51	99.49	90.86	99.25	99.24	88.63
(10, 5)	-0.1	13.12	13.05	8.64	7.34	7.29	5.61	1.99	2.00	2.21
	-0.3	39.74	39.55	23.76	18.86	18.78	13.31	11.19	11.29	8.79
	-0.9	99.95	99.95	96.15	99.41	99.39	86.77	98.68	98.69	81.45

Note Nominal size is 5%, and sample size  $T = 100$

In spite of the first case, calculated power values under the second case clearly underline differences between the new combined test statistics and  $s_{\alpha,\alpha(\beta),\alpha\beta}$ . As seen in Table 4,  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{2T}$  have considerably more power than  $s_{\alpha,\alpha(\beta),\alpha\beta}$ . These results indicate that the  $R_{1T}$  and  $R_{2T}$  test statistics approach of Caner and Hansen (2001) could be able to deal with partial stationary condition.

According to results in Table 5, which are calculated under the asymmetric stationary error case, it is state that  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{2T}$  test statistics have significantly better power than  $s_{\alpha,\alpha(\beta),\alpha\beta}$  tests. All simulated power values indicate that in the presence of nonlinear asymmetric error process, the  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{2T}$  test statistics have sufficient power relative to Leybourne et al. (1998) tests. Additionally, power values of all test statistics decline with decreasing  $\rho$  values. Also, increasing speed transition causes power loss for all test statistics under the first and third case. However, changing the value of  $\gamma$  does not have notable effect on the power values of  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{1T}$  and  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{2T}$  under the partial stationary case.

### 4 Empirical Application

Because the impact of the 2008 crisis can be easily observed, the monthly industrial production index series of Turkey and the USA are chosen for this empirical study.<sup>1</sup> The raw seasonally adjusted time series are collected from the World Bank’s Statistical Capacity Indicators database over the period of January 2004 to March 2020. The standard ADF statistics,  $s_{\alpha,\alpha(\beta),\alpha\beta}$  statistics of Leybourne et al. (1998),  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{1T}$ , and  ${}^{\alpha,\alpha(\beta),\alpha\beta}R_{2T}$  statistics are calculated for natural logarithm of industrial production index series. Before computation of the test statistics, estimated parameters of the smooth transition models (11), (12), and (13) are given in Table 6.

**Table 6** Estimated parameters of smooth transition

	Model A		Model B		Model C	
	Turkey	USA	Turkey	USA	Turkey	USA
$\delta_1$	3.925	4.543	3.987	4.535	3.927	4.518
$\delta_2$	0.964	0.079	-0.120	-0.137	-0.043	-0.116
$\varphi_1$	-	-	0.005	0.001	0.007	0.002
$\varphi_2$	-	-	-	-	-0.002	-0.001
$\gamma$	0.021	0.136	1.669	0.767	0.738	0.629
$\tau$	0.511	0.583	0.291	0.291	0.280	0.288
Mid-point date	May 2012	Jul 2013	Oct 2008	Oct 2008	Aug 2008	Sep 2008

<sup>1</sup> Data and R programming language codes that replicate the empirical study of this paper are available in author’s personal GitHub page <https://github.com/mehmet-ozcan/2021-bookchapter-1>.

The first question to ask is whether the models could estimate the correct mid-point dates. According to Table 6, the estimated parameters of Model B and Model C could identify the 2008 global economic crisis. However, mid-point dates estimated with Model A indicate different time points. The natural logarithm industrial production of Turkey and the USA with their smooth transitions is presented in Figs. 1 and 2, respectively. The time path of the estimated smooth transition of Model A for both countries miss a noticeable structural change in 2008.

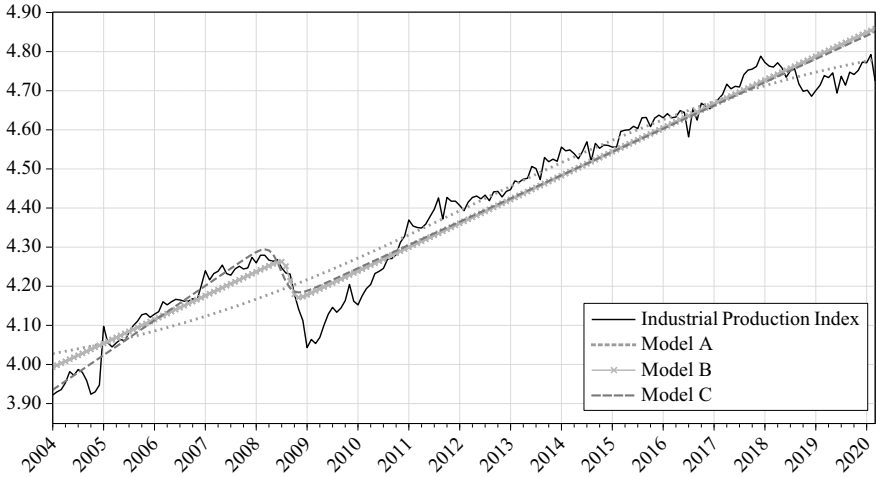


Fig. 1 Natural logarithm industrial production index of Turkey and fitted smooth transitions

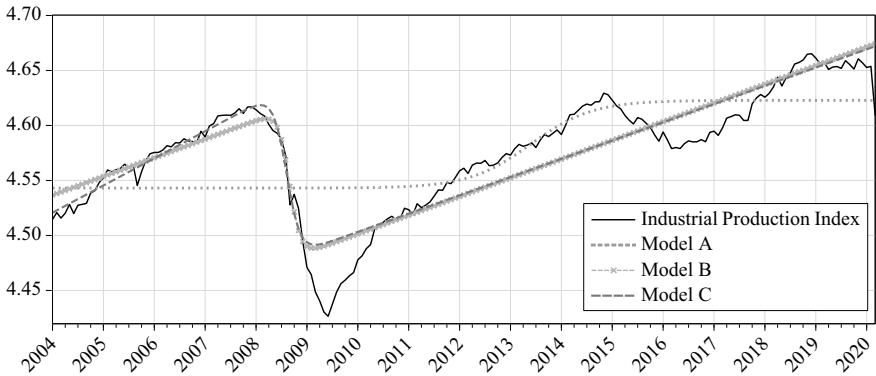


Fig. 2 Natural logarithm industrial production index of USA and fitted smooth transitions

**Table 7** Caner and Hansen (2001) threshold effect Wald test statistics ( $W_T$ )

	Turkey			USA		
Residuals	$\hat{\nu}_t^{Model A}$	$\hat{\nu}_t^{Model B}$	$\hat{\nu}_t^{Model C}$	$\hat{\nu}_t^{Model A}$	$\hat{\nu}_t^{Model B}$	$\hat{\nu}_t^{Model C}$
$W_T$	32.524	27.595	10.345	36.195	24.204	11.031
p-value	0.073	0.198	0.134	0.006	0.026	0.403

Note p-values are calculated from 1000 bootstrap replication

Another essential point is the estimated values of speed of transition parameter  $\gamma$ . Especially for Turkey, the smooth transition of Model A could not detect any structural change. Therefore, the value of estimated  $\gamma$  is close to zero. In contrast, the  $\gamma$  value estimations from Model B and Model C state a structural change. In addition, the estimated speed of transition values of Turkey are greater than  $\gamma$  estimations for the USA.

As suggested in Caner and Hansen (2001), before calculation of the  $R_{1T}$  and  $R_{2T}$  test statistics, the time series should be tested against the null hypothesis of  $H_0 : \theta_1 = \theta_2$ . This investigation is done to be sure of the nonlinear threshold autoregressive structure of the residual series of smooth transition models. For this purpose, the Wald statistics ( $W_T$ ), which was introduced by Caner and Hansen (2001), is used. Calculated results are shown in Table 7.

Findings from Table 7 state that the null of no threshold effect can be rejected for the residual series of Model A for Turkey and residual series of Model A and Model B for the USA. As mentioned before, Model A does not offer proper smooth transition for noticeable structural change in 2008. It is clear that the fitted smooth transition from Model B for the industrial production index of the USA could estimate the correct structural change date, and its residual series has significant threshold effect. In other words, the industrial production index of the USA contains the characteristics of both structural change and regime switching. Thus,  ${}^{\alpha(\beta)}R_{1T}$  and  ${}^{\alpha(\beta)}R_{2T}$  are the most appropriate test statistics to test the unit root null ( $H_0 : \rho_1 = \rho_2 = 0$ ) for the industrial production index of the USA. Calculated standard linear ADF test statistics ( $\tau_\tau$ ),  $s_{\alpha, \alpha(\beta), \alpha\beta}$  statistics of Leybourne et al. (1998), and  ${}^{\alpha, \alpha(\beta), \alpha\beta}R_{1T}$  and  ${}^{\alpha, \alpha(\beta), \alpha\beta}R_{2T}$  statistics are reported in Table 8.

Calculated test statistics indicate that the unit root null hypothesis could be rejected by only  ${}^\alpha R_{1T}$  and  ${}^\alpha R_{2T}$  for the series of both countries. However, it is obvious that the proper break date could not be estimated with Model A, and the calculated residual series of Model A does not have a statistically significant threshold effect. For this reason, rejection of the null hypothesis for  ${}^\alpha R_{1T}$  and  ${}^\alpha R_{2T}$  is not reliable. As other test results state, it is a more reliable way to conclude that the null hypothesis of unit the root cannot be rejected for the industrial production index series of both countries.

**Table 8** Empirical application of unit root tests to industrial production index of Turkey and USA

	$\tau_\tau$	$s_\alpha$	$s_{\alpha(\beta)}$	$s_{\alpha\beta}$	$\alpha R_{1T}$	$\alpha R_{2T}$	$\alpha^{(\beta)} R_{1T}$	$\alpha^{(\beta)} R_{2T}$	$\alpha\beta R_{1T}$	$\alpha\beta R_{2T}$
Turkey	-2.162	-2.655	-2.201	-2.272	33.465*	33.465*	21.596	21.596	12.468	12.524
USA	-2.848	-3.273	-1.353	-1.381	39.350*	39.350*	8.240	8.240	8.141	8.141

*Note* Natural logarithm of the series are tested. The superscript \* indicates significance at 1% according to critical values for  $T = 200$

## 5 Conclusion

This paper described a new unit root testing procedure that combines the smooth transition process of LNV and the threshold unit root tests of Caner and Hansen (2001). The empirical size values are computed under different DGP cases, and the Monte Carlo simulation results indicate that it is difficult to observe noticeable size distortions. The Empirical power values of the new combined tests are also simulated under different DGP cases and compared with the simulated power values of the unit root tests of Leybourne et al. (1998). For a linear symmetric stationary error term, there is no remarkable difference between the power values of the two tests. However, the new combined unit root test has more power than the unit root tests of Leybourne et al. (1998) in the case of a nonlinear asymmetric stationary error term and the case of a nonlinear asymmetric partial stationary error term. The new unit root testing process is illustrated by an application on the industrial production indices of Turkey and the USA. This application also includes a linear ADF test and Leybourne et al. (1998)'s tests. The results of empirical application state that there is strong evidence to indicate that the industrial production index series of the USA contains effects of both structural change and regime switching dynamics. Under this condition, none of the calculated test statistics can reject the null hypothesis of unit root for both series.

**Acknowledgements** This study is based on a PhD thesis by Mehmet Özcan (Özcan 2019) done under the supervision of Professor Funda Yurdakul.

## References

- Caner M, Hansen BE (2001) Threshold autoregression with a unit root. *Econometrica* 69(6):1555–1596
- Dickey DA, Fuller WA (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 1057–1072
- Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc* 74(366a):427–431
- Enders W, Granger CWJ (1998) Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates. *J Bus Econ Stat* 16(3):304–311
- Leybourne S, Newbold P, Vougas D (1998) Unit roots and smooth transitions. *J Time Ser Anal* 19(1):83–97
- Özcan M (2019) Yapısal Kırılma Altında Eşik Birim Kök Testlerinin İncelenmesi. Doctoral thesis, Gazi University, Ankara, Turkey. <https://tez.yok.gov.tr/UlusalTezMerkezi/TezGoster?key=vjszP7PzV0HebcjFEvDfwH8NiSvnr2XbO7qFmKNjHeMNbwErLJNK5ZITfjlXn12>
- Perron P (1989) The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* 1361–1401
- Petrucelli JD, Woolford SW (1984) A threshold AR (1) model. *J Appl Probab* 270–286
- Sollis R (2004) Asymmetric adjustment and smooth transitions: a combination of some unit root tests. *J Time Ser Anal* 25(3):409–417

- Tong H (1978) On a threshold model. In: Chen CH (ed) Pattern recognition and signal processing. Sijhoff and Noordhoff, Amsterdam, pp 575–586
- Vougas DV (2006) On unit root testing with smooth transitions. *Comput Stat Data Anal* 51(2):797–800
- Zivot E, Andrews DWK (1992) Further evidence on the great crash, the oil price shock and the unit root hypothesis. *J Bus Econ Stat* 10:251–270

# Jump Connectedness in the European Foreign Exchange Market



Emawtee Bissoondoyal-Bheenick, Robert Brooks, and Hung Xuan Do

**Abstract** We assess the jump connectedness (spillover) among five Group-of-Ten European currencies, namely the Swiss Franc, the Euro, the British pound, the Norwegian Krone, and the Swedish Krone. Our analysis covers a period starting from January 1999 to January 2018. Overall, we find evidence of jump connectedness in the Group-of-Ten European currencies, in which the Euro is the largest net transmitter and the British pound is the largest receiver of jump connectedness. Jump connectedness between the Euro and the Swiss Franc is the strongest followed by the Euro-Norwegian Krone and the Swiss Franc-Swedish Krone pair. Total jump connectedness among the five Group-of-Ten European currencies is time-varying and sensitive to the extreme events such as the Eurozone Debt Crisis. However, the good news is that their jump connectedness is in a downward trend, declining about a half of its peak observed in early 2007.

**Keywords** Jump connectedness · BNS G jump statistics · Foreign exchange · Europe

## 1 Introduction

Significant discontinuities occasionally exist in the financial returns. These are regarded as jumps relative to the smooth component of the return volatility. A body of the literature has documented the presence of jumps in different asset returns as well as their critical roles in financial management (see Duffie et al. 2000; Eraker et al.

---

E. Bissoondoyal-Bheenick

School of Economics, Finance and Marketing, RMIT University, Melbourne, Australia  
e-mail: [banita.bissoondoyal-bheenick@rmit.edu.au](mailto:banita.bissoondoyal-bheenick@rmit.edu.au)

R. Brooks (✉)

Department of Econometrics and Business Statistics, Monash University, Clayton, Australia  
e-mail: [Robert.brooks@monash.edu.au](mailto:Robert.brooks@monash.edu.au)

H. X. Do

School of Economics and Finance, Massey University, Palmerston North, New Zealand  
e-mail: [h.do@massey.ac.nz](mailto:h.do@massey.ac.nz)



2003; Johannes 2004; Huang and Tauchen 2005; Lee and Mykland 2008; Pukthuanthong and Roll 2015). Statistically, the existence of jumps helps in explaining the fat tails (excess kurtosis) and asymmetry (skewness) level of the financial return distributions. In finance, a recognition of jumps may improve the quality of financial activities such as asset pricing (e.g., see Piazzesi 2003, for bond; Duffie et al. 2000, for options on bond, currencies, and equities) and financial risk management (e.g., see Lee and Mykland 2008, for derivative hedging; Zhou et al. 2019, for optimal portfolio allocation).

Jumps might appear for many reasons. In general, large changes link with jumps. Particularly, extreme events such as political changes, terrorist attacks, financial crises and insolvency news, or shocks to energy prices can cause such changes (see Pukthuanthong and Roll 2015). Huang and Tauchen (2005) shows that jumps accounts for around 4.5 to 7% of the daily variance of the S&P index including cash index and index futures. Given the evidence on the presence of jumps and their crucial roles in financial activities, one of the most important questions regarding risk management is whether jumps transmit across markets.

The objective of this chapter is to assess the existence of jumps and jump spillover (or connectedness) in the five G10 European currencies, including Euro (EUR), British pound (GBP), Swiss franc (CHF), Norwegian krone (NOK), and Swedish krona (SEK). Cross-country jump analysis is beneficial for global investors and financial risk managers who hedge risk by international portfolio diversification. If jump connectedness among markets is high, international diversification would be strongly and adversely affected by extreme events that happened in one market. Few studies have focused on the jump connectedness in the stock markets (e.g., Asgharian and Bengtsson 2006; Bengtsson 2006; Asgharian and Nossman 2011; Jawadi et al. 2015). However, there is no study investigating a similar issue in the currency markets. Given the importance of the G10 European currencies as discussed later, the question on existence of their jump connectedness deserves significant attention. Hence our key contribution is to extend the literature on jump connectedness to the European foreign exchange (FX) market, and we show that jumps and jump connectedness do exist in the five G10 European currencies. These linkages are statistically significant over the sample period from January 1999 to January 2018. Our dynamic analysis illustrates that the jump connectedness among the five G10 European currencies is time-varying and responsive to the extreme events such as the Eurozone Debt Crisis. However, the good news is that their jump connectedness is in a decreasing trend, whose strength (as measured by the connectedness index) has declined about a half from its peak in 2007.

The importance of the currency market firstly is due to the large volume of transactions on a daily basis. This is one of the biggest markets in the world, offering high risk and high returns to many individual and institutional investors.<sup>1</sup> The currency market is tied to global markets and significantly contributes to the interconnectedness of markets. With globalization, trading in the G10 currencies have increased

---

<sup>1</sup> See: <https://www.rba.gov.au/publications/bulletin/2016/dec/pdf/rba-bulletin-2016-12-developments-in-foreign-exchange-and-otc-derivatives-markets.pdf>:

considerably during the past years. Kitamura (2010) shows that the FX market has a high degree of integration, in particular for the most tradeable currencies. The monetary policies of countries are related closely to the FX volatility in the market, see, for example, Devereux and Engel (2003).

Among G10 currencies, we focus on jump connectedness in European markets due to several reasons. Firstly, analyzing the European currencies is important following several critical events including the European debt crisis and Brexit, given that there have been ongoing policy changes which has an impact on the European economies and the fluctuation and jumps of the currencies values. Following the European debt crisis, the European Central Bank (ECB) has kept interest rates at negative levels with a view to reduce the supply of money to stimulate the economy and increase the growth level. When interest rates are very low or negative, it implies that the EURO is more expensive as compared to the other trading partners. Since its inception, the EUR has had significant fluctuation and been quite volatile as compared to other major currencies. Its lowest point has been in point in 2000, and it peaked its value in 2008. The Global Financial Crisis in the U.S. (GFC) has definitely an impact on the value of the EUR, but it later bounced backed and in November 2011, the EUR traded almost as high as it was at before the GFC in 2007. In addition to the Eurozone debt crisis, the global markets as well as the European markets have been volatile because of Brexit in 2016, when the decision of Britain to leave the European Union (EU). Secondly, while the EUR is being used in 19 European countries (2018), there are a number of closely related countries in terms of trade connectedness and geographical proximity that do not use the EUR as their currency. Among the G10 currencies, these include Switzerland (Swiss Franc (CHF)), Great Britain (British pound (GBP)), Norway (Norwegian Krone (NOK)), and Sweden (Swedish Krone (SEK)).

Our study contributes to the literature in two main aspects: (1) we add new evidence on existence of jumps in the five G10 European currencies and (2) we provide a comprehensive study about the jump connectedness among these currencies in terms of both static and dynamic analyses. Our analyses show that jump connectedness exists in the top five European currencies. Among them, we find that EUR is the largest net transmitter of the jump connectedness (EUR is the most dominant currency) while GBP is the largest net receiver. Regarding pairwise analyses, the EUR-CHF pair shows the strongest jump connectedness followed by the EUR-NOK and the CHF-SEK pair. Finally, we find that the jump connectedness among the five G10 European currencies changes over time and it is sensitive to the extreme events such as the crises. However, global investors may still have diversification benefit from these currencies because their jump connectedness has decreased nearly a half since 2007.

The remainder is organized as follows. Section 2 details the construction of jumps and econometrics frameworks, and Sect. 3 describes the data and existence of jumps in the five G10 European currencies. We discuss the jump connectedness results in Sect. 4 and provide our conclusion in Sect. 5.

## 2 Construction of Jumps and Methodology

### 2.1 Construction of Jumps

Due to an unavailability of the high frequency data for the five G10 European currencies, we follow Pukthuanthong and Roll (2015) approach to use the daily data for constructing the monthly Barndorff-Nielsen and Shephard (2006) G jump statistics (BNS-G, hereafter). Under the no jump hypothesis, the BNS-G follows a standard normal distribution asymptotically. However, if there is at least one jump in a month, BNS-G tends to be negative. Pukthuanthong and Roll (2015) have considered alternative jumps and found that the test power of BNS-G is favorable in comparison with other methods. This result adds evidence in explaining why BNS-G is regarded as the most prominent jump detection. Therefore, we employ the BNS-G as our measure of the jump statistics. The monthly BNS-G for currency  $f$  in month  $m$  ( $G_{f,m}$ ) can be calculated based on the monthly bipower variation ( $B_{f,m}$ ), monthly squared variation ( $S_{f,m}$ ), and monthly quarticity ( $Q_{f,m}$ ) as follows:

$$G_{f,m} = \frac{\frac{\pi}{2} B_{f,m} - S_{f,m}}{\sqrt{\frac{\vartheta \pi^2}{4} Q_{f,m}}}, \quad (1)$$

where  $\vartheta = (\pi^2/4) + \pi - 5$ . The monthly bipower variation, monthly squared variation, and monthly quarticity for currency  $f$  in month  $m$  can be constructed respectively using the daily log return as follows:

$$B_{f,m} = \frac{1}{D_m - 1} \sum_{d=2}^{D_m} |r_{f,d,m}| |r_{f,d-1,m}|, \quad (2)$$

$$S_{f,m} = \frac{1}{D_m} \sum_{d=1}^{D_m} (r_{f,d,m})^2, \quad (3)$$

$$Q_{f,m} = \frac{1}{D_m - 3} \sum_{d=4}^{D_m} |r_{f,d,m}| |r_{f,d-1,m}| |r_{f,d-2,m}| |r_{f,d-3,m}|, \quad (4)$$

where  $r_{f,d,m}$  indicates the log return of currency  $f$  on day  $d$  in month  $m$ .  $D_m$  is the number of days in month  $m$ .

## 2.2 The Vector Autoregressive Model and Impulse Response Function

We investigate the spillover effect between selected European currencies' jumps by constructing the generalized impulse response function (see Pesaran and Shin 1998) as well as the dynamic generalized spillover indices (see Diebold and Yilmaz 2012) within a Vector Autoregressive (VAR) model. As we focus our analyses on five European FX currencies (EUR, GBP, NOK, SEK, and CHF), the generalized approach of the impulse response and spillover index is preferred because there is no clear economic intuition and evidence in ordering the instantaneous causality among the chosen currencies. In this way, we can obtain a unique set of results regardless of alternative orderings in the VAR system. The interpretation for the generalized impulse response analysis of BNS-G measure is straightforward. Spillover effect exists between jumps of two currencies if one's BNS-G statistic positively respond to an exogenous shock to another's BNS-G statistic.

We consider a 5-dimensional vector  $G_t$  that collects the monthly BNS-G measures of the five European currencies over time  $t$ ,  $G_t = (G_{EUR,t}, G_{GBP,t}, G_{NOK,t}, G_{SEK,t}, G_{CHF,t})'$ . A VAR model of  $G_t$  can be specified as,

$$G_t = \sum_{i=1}^p A_i G_{t-i} + e_t, \quad t = 1, 2, \dots, T, \quad (5)$$

where  $e_t \sim (0, \Sigma_e)$  is an identically and independently distributed error term.  $\Sigma_e = \{\sigma_{ij}; i, j = 1, \dots, 5\}$  denotes the variance-covariance matrix of  $e_t$ .  $A_i$  is the  $(5 \times 5)$  coefficient matrix corresponding to the lag  $i$  of  $G_t$ .  $p$  is the lag length of the VAR model which is determined based on the Akaike Information Criterion. If all the roots of the lag polynomial,  $|A(z)| = |I_K - \sum_{i=1}^p A_i z^i| = 0$ , fall outside the unit circle, the VAR model is covariance stationary and it can be rewritten in its moving average representation as follows:

$$G_t = \sum_{i=1}^{\infty} \Phi_i e_{t-i} \quad (6)$$

where  $\Phi_i$  can be calculated recursively as,  $\Phi_i = \sum_{j=1}^p \Phi_{i-j} A_j$  with  $\Phi_0$  is the  $(5 \times 5)$  identity matrix.

Pesaran and Shin (1998) shows that the matrix of the generalized impulse response at the horizon  $h$  (with one standard deviation as the unit shock) can be computed as

$$GIRF_h = \Phi_h \Sigma_e \Lambda \quad h = 0, 1, 2, \dots \quad (7)$$

in which,  $\Lambda$  is a  $(5 \times 5)$  diagonal matrix characterized by the standard deviation of  $e_t$ ,  $\Lambda = \text{diag} \left\{ \frac{1}{\sqrt{\sigma_{11}}}, \frac{1}{\sqrt{\sigma_{22}}}, \dots, \frac{1}{\sqrt{\sigma_{55}}} \right\}$ .

### 2.3 The Generalized Connectedness Index

The generalized forecast-error variance decomposition (GFEVD) at horizon  $H$  within our VAR model can be defined as follows:

$$\omega_{ij}^g(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (u_i' \Phi_h \Sigma_e u_j)^2}{\sum_{h=0}^{H-1} (u_i' \Phi_h \Sigma_e \Phi_h' u_i)}, \quad (8)$$

where  $u_i$  denotes a  $(5 \times 1)$  vector, in which its  $i$ th element is one and zeros elsewhere.

Diebold and Yilmaz (2012) suggest that the generalized connectedness indices can be constructed using the normalized GFEVD. The total connectedness index represents the spillovers contribution from all the currencies to the total forecast error variance,

$$S^g(H) = \frac{\sum_{i,j=1, i \neq j}^5 \tilde{\omega}_{ij}^g(H)}{5} \times 100, \quad (9)$$

where  $\tilde{\omega}_{ij}^g$  is the normalized GFEVD that can be calculated as

$$\tilde{\omega}_{ij}^g(H) = \frac{\omega_{ij}^g(H)}{\sum_{j=1}^K \omega_{ij}^g(H)}, \quad (10)$$

with a note that  $\sum_{j=1}^5 \tilde{\omega}_{ij}^g(H) = 1$  and  $\sum_{i,j=1}^5 \tilde{\omega}_{ij}^g(H) = 5$  by construction.

The directional connectedness received by a currency  $i$  from four remaining currencies are constructed as follows:

$$S_{i \leftarrow}^g(H) = \frac{\sum_{j=1, j \neq i}^5 \tilde{\omega}_{ij}^g(H)}{\sum_{i,j=1}^5 \tilde{\omega}_{ij}^g(H)} \times 100. \quad (11)$$

Similarly, the directional connectedness from a currency  $i$  to four remaining currencies is computed as follows:

$$S_{i \rightarrow}^g(H) = \frac{\sum_{j=1, j \neq i}^5 \tilde{\omega}_{ji}^g(H)}{\sum_{i,j=1}^5 \tilde{\omega}_{ij}^g(H)} \times 100. \quad (12)$$

As a result, net connectedness transmitted by the currency  $i$  to four remaining currencies is

$$S_i^g(H) = S_{i \rightarrow}^g(H) - S_{i \leftarrow}^g(H). \quad (13)$$

If  $S_i^g(H) > 0$ , we call currency  $i$  is a net transmitter of the connectedness, whereas, if  $S_i^g(H) < 0$ , it is the net receiver.

### 3 Existence of Jumps in Five G10 European Currencies

We collect daily spot exchange rates for the five G10 European currencies from Datastream, including EUR, GBP, NOK, SEK and CHF from 1st January 1999 to 31st January 2018, comprising of 4,979 daily observations. We calculate the daily returns for each currency as the difference between logarithmic exchange rates. The daily returns are then employed to construct the monthly BNS-G measures following Pukthuanthong and Roll (2015) as described in the Sect. 2.1. Therefore, we finally have 229 monthly observations for each currency's BNS-G. In the next section, we discuss the characteristics of the BNS-G measures in the five G10 European currencies.

We provide the summary statistics for the monthly BNS-G statistics of the five currencies in Table 1. As discussed earlier in Sect. 2.1, BNS-G would follow a standard normal distribution if there is no jump. However, when there are jump(s) in a month, BNS-G would be negative. As such, the mean, standard deviation, skewness, and kurtosis of the BNS-G should be close to 0, 1, 0, and 3, respectively under the hypothesis of no jump.

Table 1 shows that means of all BNS-Gs are statistically significantly negative, indicating existence of jumps in all currencies. The mean of BNS-G for GBP is statistically significant at  $-0.312$ , NOK at  $-0.328$ , SEK at  $-0.364$ , CHF at  $-0.395$ , and EUR at  $-0.437$ . The negative means across all five currencies indicate the presence of multiple jumps in a month for each of the currencies. The standard deviation of BNS-G for GBP is 0.640, NOK is 0.651, SEK is 0.840, CHF is 0.717,

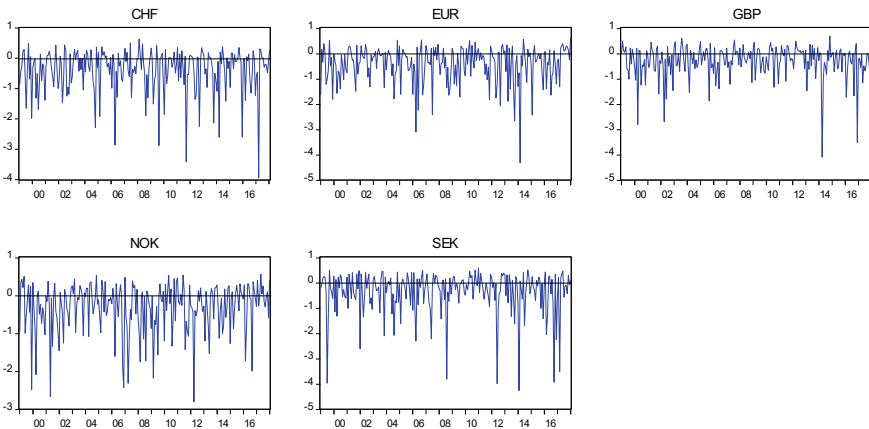
**Table 1** Summary statistics of BNS-G measures of the five G10 European currencies. This table reports the summary statistics and three unit root test statistics for the monthly BNS-G-jumps of each European currency under consideration. The asterisks \*\*\*, \*\*, \* indicate the null hypothesis ( $H_0$ ) is rejected at 1%, 5%, and 10% level of significance, respectively. Under  $H_0$  of Jarque–Bera test, the series is normally distributed. Regarding Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) test, the series is nonstationary under the  $H_0$ . In Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test,  $H_0$  refers to when the series is stationary

	CHF	EUR	GBP	NOK	SEK
Mean	−0.395	−0.437	−0.312	−0.328	−0.364
Median	−0.185	−0.256	−0.169	−0.161	−0.130
Maximum	0.623	0.582	0.680	0.556	0.573
Minimum	−3.945	−4.338	−4.106	−2.817	−4.274
Std. Dev.	0.717	0.724	0.640	0.651	0.840
Skewness	−1.860	−1.587	−2.272	−1.411	−2.353
Kurtosis	7.497	6.858	11.435	5.068	9.865
Jarque–Bera	325.0***	238.1***	876.0***	116.8***	660.9***
ADF	−15.46***	−15.18***	−16.04***	−14.73***	−15.69***
PP	−15.73***	−15.18***	−16.03***	−14.73***	−15.73***
KPSS	0.039	0.063	0.116	0.205	0.074

and EUR is 0.724. These figures are all smaller than 1 indicating that the distribution of all five BNS-G is more concentrated around the central location than the standard normal distribution. Together with high values of kurtosis (greater than 3) observed in all five BNS-Gs, it implies that all five BNS-Gs exhibit leptokurtic distributions with higher peaks and fatter tails than a normal distribution. Skewness of BNS-G is negative for all five currencies ( $-2.272$  for GBP,  $-2.353$  for SEK,  $-1.411$  for NOK,  $-1.86$  for CHF, and  $-1.587$  for EUR). These indicate that during some months in our sample, jump statistics are significantly smaller than the expected value under the null hypothesis of no jump. Hence, there is more chance of jumps within the month. Both the reported skewness and kurtosis show significant departure from asymptotic normality (see Table 1).

The Jarque–Bera tests further support the non-normal distribution of BNS-Gs in all cases, which confirms the existence of jumps in all currencies under consideration. The individual minimum and maximum also confirm a strongly negativity of the BNS-G values. Across all the currencies, the minimum is significantly negative, and the maximum is less than one. All three unit-root tests confirm the stationarity of all BNS-Gs, indicating the suitability of applying a short-memory multivariate modeling technique such as our VAR framework as discussed in Sect. 2.

From Table 1, we have established the significant presence of jumps in all the currencies that we have in the sample. We further plot the time series of BNS-Gs of five European currencies in Fig. 1. This figure further supports Table 1 on the existence of jumps in five selected European currencies. We frequently observe the BNS-Gs in the negative region with many highly negative values (recall that negativity of BNS-G shows there is one or more jumps within a month). Pukthuanthong and Roll (2015) indicate that jumps can be due to a number of factors including: unexpected changes in the return distribution’s characteristics, extreme events, or exogenous shocks to primary global factors. In our study, the sample period is long



**Fig. 1** Selected European currencies’ jumps plots. This figure shows time series plot of BNS G-jump statistics of five selected European currencies

enough to cover significant events in the global FX markets which has potentially contributed to the jumps in each currency.

Some of the events in our sample period from 1999 to 2018 are worth noting. The first example of a major event is the European Debt Crisis from 2009 to 2013. There have been a number of countries which required bail outs of debt, including Greece, Portugal, and Spain among others, leading to economic recession and downturns. The debt crisis required the need for new which changed Eurozone's financial system. The major outcome of the debt crisis has been the impact on the EUR and some of major currencies in Europe which were subject to volatility and call in value in the wake of this crisis, and it is still a factor in the EUR's value to date.

Our sample also includes the Greek debt crisis from 2010 to 2017. The Greek economy was among the most severely affected Eurozone crisis. Several rescue measures and policies were introduced to stabilize the economy. One of the consequences of the Greek crisis is the impact that it has had on the periphery countries (both the stock and currency market). There was also spillover effect from the Greek crisis to the value of the Swiss Franc, CHF, and the property market in 2015. Investors view the CHF as a safe haven given it is one of few independent European currencies and hence with the Greek crisis, there was a flight to quality and investors invested in the CHF largely.

Another event includes the Russian Financial Crisis from 2014 to 2015, which originated from the collapse of Russia's currency—the Rouble. While the main reasons for the collapse in the Rouble are decreasing oil prices and economic sanctions, these have a direct impact on the value of currencies in global market.

Finally, our sample includes Brexit in 2016. While Britain voted to leave the EU in 2016, the discussion in the lead up to Brexit created volatility in the global FX market. Following the vote in Britain, the GBP/EUR rate has dropped by 7.5% in a week, and this depreciation went further down to 20% over a period of only two weeks.<sup>2</sup> All these significant major events in our sample period have had significant impacts on the global currencies and in particular the G10 currencies, which partly explains the presence of significant jumps in the G10 European currencies in our sample. In the next section, we discuss empirical results regarding the existence and dynamic behavior of jump connectedness in European currency markets.

## 4 Jump Connectedness in the Five G10 European Currencies

Table 1 clearly confirms that we have the presence of significant jumps in each of the currencies that we are considering in this study. Hence, our next step is to assess the connectedness in the jumps of these currencies. We therefore employ a VAR model with the Generalized Impulse Response Function (see Pesaran and Shin 1998) and

---

<sup>2</sup> <https://moneytransfercomparison.com/major-economical-events-which-impacted-currency-rates/>.



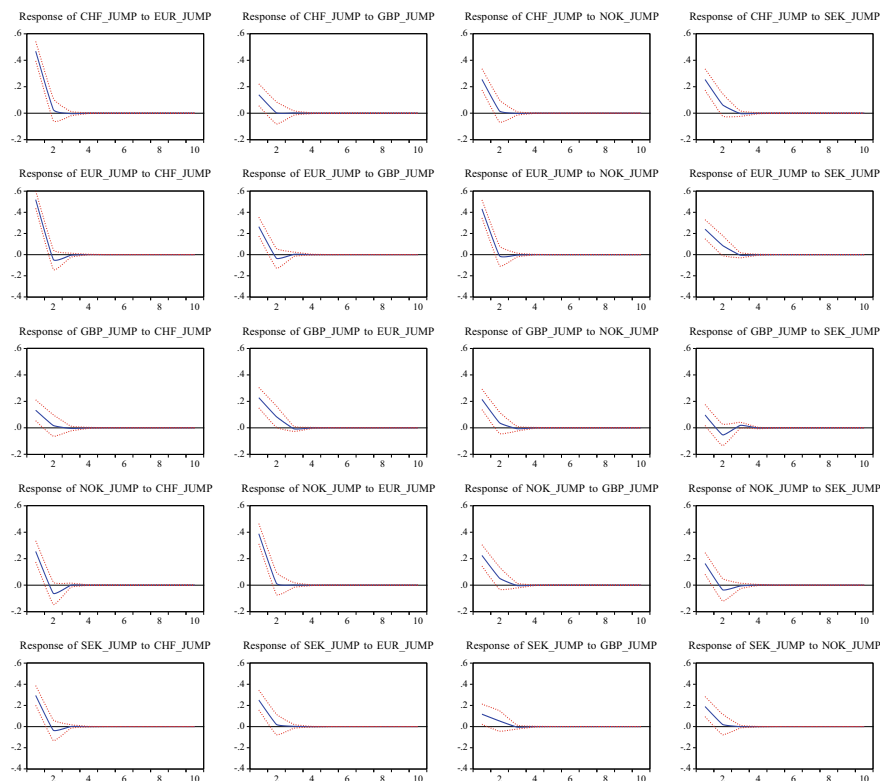
generalized connectedness index (see Diebold and Yilmaz 2012, 2015) to investigate the jump connectedness among selected European currencies. Conclusion about the stationarity of all BNS-G measures as discussed in previous section supports our application of a VAR model in analyzing the interrelationship among currency jumps. Using the Akaike Information Criterion, we find an optimal lag length of 1 for our VAR model. Our model diagnostics via the auto correlation function and Ljung–Box test show that the error terms mimic the white noise processes. In addition, we find that all roots of our estimated VAR model fall outside the unit circle, indicating that our VAR model is covariance stationary. Therefore, the generalized impulse response function and connectedness index approach can be adequately applied to investigate the jump connectedness.

#### 4.1 *Pairwise Jump Connectedness*

As explained earlier, if there is connectedness (or spillover) among jumps of different currencies, there will be positive relationship between their BNS-Gs. That is, we would find a positive response of one BNS-G to an exogenous shock to another BNS-G in the impulse response analysis.

We plot the generalized impulse response of one currency's BNS-G to a one standard deviation exogenous shock to another currency's BNS-G in Fig. 2. Figure 2 shows statistically significant and positive impulse response among jumps of five selected European currencies. The positive generalized impulse response indicates a positive spillover effect between currencies' jumps, that is, if jumps occurs in one currency due to an exogenous shock, it tends to spread out and connects to occurrence of jumps in other currencies in the system. We find that the spillover effect generally lasts until the second month.

Table 2 reports the results of the static jumps spillover indices in the European currencies over the whole sample period. The indices are calculated from the GFEVD based on 12-step-ahead forecasts (i.e., one year). The GFEVD is based on a five-dimension VAR with a lag length of 1. The diagonal of this table measures the own currency jump spillover, while off diagonals measure the cross-currency jump spillover between two currencies. The rows are the contribution "to" other currencies while the column is the contribution "from" currencies. The difference between sum of contribution to other currencies and sum of contribution from other currencies is the net spillover from one currency to all other currencies. This table shows the static total and also the directional and net spillover among the currency pairs. We find interesting results from Table 2. The total connectedness index indicates an average contribution of 35.1% in the total jumps of the currencies. As expected, the own currency jump spillover represented by the diagonal elements explains the largest proportion of the forecast errors. GBP is at 77%, followed very closely by SEK at 76.7%, then we have NOK at 62.4%, CHF at 57.3%, and finally EUR at 50.9% for the own currency jump.



**Fig. 2** Generalized impulse response function between jumps of European currencies. This figure shows the generalized impulse response of one currency’s G-jump statistic to a one standard deviation exogenous shock to another currency’s G-jump statistic. Positive generalized impulse response indicates positive spillover effect between currencies’ jumps. That is, if jump(s) occurs in one currency due to an exogeneous shock, it tends to spread out and connects to occurrence of jump(s) in other currencies in the system

While the own currency jump spillover of EUR is lowest, it is the largest contributor of jump spillover to the other currencies. The EUR contributes 61.1% of spillover effect to other currencies, while it receives 49.1% of that from them. The second largest contributor for the jump spillover is the CHF which contributes 44.1% but receives 42.2% from other currencies. In fact, the CHF is the most stable in terms of receiving and transmitting spillover with the small variance between these two values, which is similar to the conclusion of Barunik et al. (2017) who find that the CHF is the most balanced currency between the role of giver and receiver of the volatility spillover effect.<sup>3</sup> The row which shows the net spillover (last row of the

<sup>3</sup> Barunik et al. (2017) investigates the asymmetric volatility spillover among six currencies, including Australian dollar (AUD), British pound (GBP), Canadian dollar (CAD), Euro (EUR), Japanese yen (JPY) and Swiss franc (CHF).

**Table 2** Overall jump spillover in five G10 European currencies using BNS-G. This table reports the jumps spillover over the whole sample period. Spillover indices are calculated from the GFEVD based on 12-step-ahead forecasts (i.e., one year). The GFEVD is based on a five-dimension VAR with a lag length of 1

	CHF	EUR	GBP	NOK	SEK	FROM
CHF	57.3	25.2	2.2	7.5	7.8	42.7
EUR	22.5	50.9	5.9	15.4	5.4	49.1
GBP	3	9.9	77	8	2.1	23
NOK	8.6	18.9	6.6	62.4	3.5	37.6
SEK	10	7.2	1.9	4.1	76.7	23.3
Contribution to others	44.1	61.1	16.6	35	18.8	Total spillover = 35.1%
Contribution including own	101.4	112	93.5	97.4	95.6	
Net spillovers	1.4	12	-6.5	-2.6	-4.4	

table) indicate that, on average, the CHF (1.4%) and EUR (12%) are the net transmitters of the jump spillover, while the GBP (-6.5%), NOK (-2.6%), and SEK (-4.4%) are the net receivers of the jump spillover. Among them, the EUR is the largest net transmitter of the spillover while the GBP is the largest net receiver.

Regarding static pairwise jump connectedness, Table 2 shows consistent results with the magnitude of generalized impulse response that we observed in Fig. 2. The BNS-Gs of EUR and CHF are most connected (spillover from EUR to CHF is 25.2% and 22.5% vice versa), following by the EUR-NOK pair (spillover from EUR to NOK is 18.9% and 15.4% vice versa), and the CHF-SEK pair (spillover from CHF to SEK is 10% and 7.8% vice versa). The lowest jump connectedness is observed in the GBP-SEK pair (with spillover from GBP to SEK at 1.9% and 2.1% vice versa).

The strongest jump connectedness between CHF and EUR can be explained by the close association of the CHF to the EUR. While worldwide financial markets were in turbulence, from the GFC in 2007 and the European Debt Crisis, the central bank of Switzerland introduced the exchange-rate peg of CHF to the EUR in 2011. Investors have always considered the CHF as a “safe haven” asset as the Swiss government is known for a balanced budget. Hence, investing in the CHF is less risky and has stable returns in the market. However, the impact of high level of investment in the CHF led to an overvaluation of the currency, which affect the economy with strong reliance on foreign trade and exports of luxury goods. To bring the CHF value down, the Swiss government implemented quantitative easing program (printing money to increase supply) and ended up with negative interest rates. With low interest rates, this implies that corporations can benefit by having loans in CHF in Europe and hence bring more pressure on the value of the CHF. In January 2015, the Swiss central bank abandoned the peg of the CHF to the EUR and among the different reasons, one of them is the fact the ECB was starting its bond buying program with a view to push down the value of the EUR. The strong association of the CHF and EUR explains the reason why for the static pairwise jump the CHF-EUR pair is the strongest.

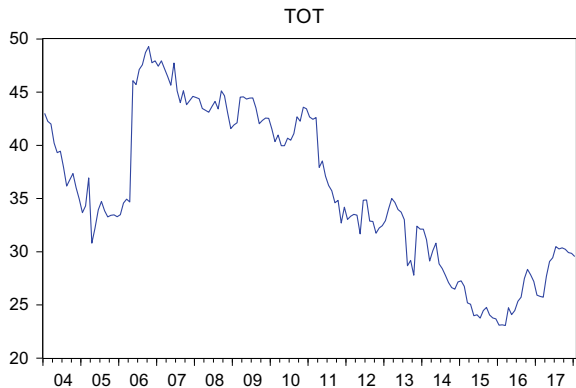
Our static jump spillover analyses further confirm the importance of the EUR in jump connectedness. The EUR is the most dominant currency in terms of the contribution to others as well as in terms of receiving jump spillover from other currencies. This finding is also supported by the net spillover values, which shows that EUR the largest transmitter of jump spillover (12%) among all the currencies. Our results are equally consistent with the literature on volatility spillover in the European FX market, see for example, Nikkinen et al. (2006), McMillan and Speight (2010), and Kitamura (2010), which find evidence on volatility spillover from EUR to GBP. In addition, Antonakakis (2012 found that EUR is the largest net transmitter of volatility spillover, whereas GBP is the largest net receiver.

### 4.2 Total Dynamic Jump Connectedness

In this section, we investigate the dynamic behavior of the total jump connectedness among the five selected European currencies. Figure 3 presents the dynamic connectedness index for all the five currencies. This index is estimated using a 60 month rolling window with a 12-step-ahead forecast horizon. The total spillover index represents the total connectedness of jump statistics of the five selected European currencies. This figure provides evidence of dynamic jump spillover and indicates that the jump spillover effect tends to decrease from early 2004 to around end of 2005 then it experienced a significant surge in 2006. After that, it is again on a decreasing trend, but the trend is more pronounced for the period from 2010 to 2016 as compared to the decreasing trend from early 2007 to 2010 approximately. From 2016, it tends to increase once again. Overall, global investors may still have diversification benefit from these currencies because their jump connectedness has decreased nearly a half since 2007.

We have already established in the previous section that the EUR is a dominant currency in our study in terms of pairwise jump connectedness. It should be noted

**Fig. 3** Total selected European currencies' jump spillover index. This figure shows the dynamic pattern of total spillover index for five selected European currencies. This index is estimated using 60 month rolling window with 12-step-ahead forecast horizon



that, while EUR was introduced in 1999, it was still on paper. EUR became in full circulation in 2002. Antonakakis (2012) indicated that the introduction of the EUR clearly altered the relative importance of other currencies in the global FX market. This impact was even more pronounced in the Eurozone for those currencies which did not use the EUR (which are all included in our sample). As more countries adopted the EUR as their currency (only 11 in 2002), the EUR gained good value and became the second most tradeable currency in the FX market in the lead up to the GFC.<sup>4</sup> The peak of jump connectedness in early 2007 shows the adoption of the EUR by other countries and the FX markets and a more stable currency as compared to the inception date.

The decreasing trend from 2007 to 2016 can be explained by three major events in the global FX market. The decreasing trend from 2007 to 2009 reflects the impact of the GFC to the Eurozone. The Eurozone was in recession in the third quarter of 2008.<sup>5</sup> The level of growth was consequently negative in the three quarters of 2008. This had an impact on all the major European FX markets. The next decline is from 2010 to 2014 which coincides with the Eurozone Debt Crisis as well the Greek Crisis. As part of the crisis measures, there has been several bailouts which had an impact on the FX markets as well as the Greek crisis brought uncertainty in the market about the Greek exit from the EU (Grexit). The negotiations between Greece and its European creditors led the financial markets in turmoil. The EUR depreciated sharply against most major currencies in 2010, including about 8% against GBP.<sup>6</sup> From 2014, the European FX markets have been volatile for two reasons: (1) the European Central Bank (ECB) embarking on the bond buying program and (2) the discussion to the lead of Brexit in 2016. The trend increases from 2016 once again as the ECB has been successful in the bond buying program and have already announced the end of the program by end of 2018.<sup>7</sup>

To further analyze this trend, we report the total directional jump connectedness effect in Fig. 4. This figure displays the directional jump spillover effect from (to) one currency to (from) all other currencies. The net spillover index shows the difference between the effects that one currency transmits and receives jump spillover from all others in the VAR system. The index is estimated using a 60 month rolling window with a 12-step-ahead forecast horizon. Panel A shows the jump spillover effect of one currency to all other four currencies, while panel B shows the sum of jump spillover effect from the four currencies to one remaining currency, and panel C is the difference between these two, that is, the net jump spillover. CHF and EUR are the net transmitters of the jump spillover effect most of the time with a note that EUR has remained as the strong net transmitter in the recent time, while CHF has mostly been the net receiver since 2012. GBP has consistently been the net receiver

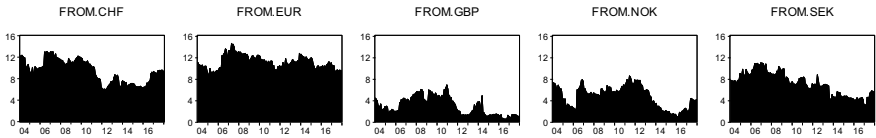
<sup>4</sup> See: <https://www.bis.org/publ/rpfx16fx.pdf>.

<sup>5</sup> See: <https://www.eubusiness.com/news-eu/1231409822.27/>.

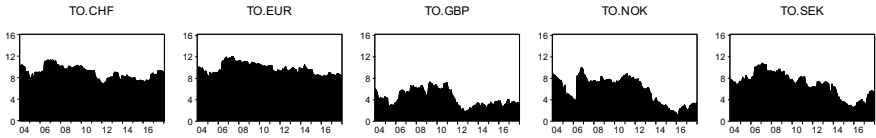
<sup>6</sup> See: <https://www.theconversation.com/how-greeces-euro-stalemate-is-hurting-currency-markets-37819>.

<sup>7</sup> See: <https://www.marketwatch.com/story/ecb-aims-to-end-bond-buying-program-by-end-of-2018-2018-06-14>.

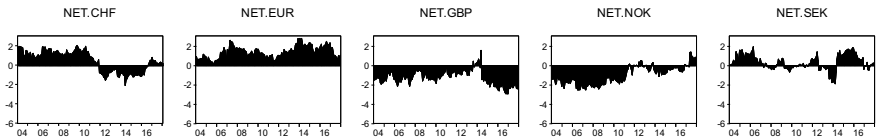
Panel A: From one jump to all other jumps



Panel B: From all other jumps to one jump



Panel C: Net jump spillover



**Fig. 4** Total directional selected European currencies’ jump spillover. This figure displays the directional spillover effect from (to) one currency’s jump statistics to (from) all other currencies’ jumps under consideration. The net spillover index shows the difference between the effects that one jump statistics transmits and receives from all other jumps statistics in the VAR system. The index is estimated using 60 month rolling window with 12-step-ahead forecast horizon. Panel A: From one jump to all other jumps, Panel B: From all other jumps to one jump, and Panel C: Net jump spillover

of the jump spillover effect. NOK has normally been the net transmitter of the jump spillover effect in the analyzed sample while SEK switches its role frequently. The general decreasing trend of the total jump spillover effect is mostly caused by the behavior of the NOK and SEK.

## 5 Conclusion

In our analysis, we extend the jump literature on the FX market by using the BNS G jump statistics to assess the existence of jumps and jump connectedness in the European FX currencies, including the CHF, EUR, GBP, NOK, and SEK for the period January 1999 to January 2018. We assess both the static and dynamic pairwise jump connectedness as well as a total jump connectedness index among selected currencies.

We analyze three key research questions, (1) do we have jumps in the European FX market by analyzing the individual markets? (2) Do we have jump connectedness

if we consider the currencies in pairs? and (3) we construct total jump connectedness index for the European FX market to assess the dynamic pattern in jump connectedness. Our overall conclusion is that jumps and jump connectedness do exist among the currencies under consideration. In particular, we find that among those selected five currencies, EUR is the largest net transmitter of the jump connectedness (EUR is the most dominant currency) while GBP is the largest net receiver. Regarding pairwise jump connectedness, we have jump connectedness between EUR and CHF is strongest followed by the EUR-NOK pair and the CHF-SEK pair. Finally, the dynamic total jump connectedness index indicates an overall declining pattern that the jump connectedness among the five G10 European currencies has decreased about a half of its peak observed in early 2007.

## References

- Anderson T, Benzoni L, Lund J (2002) An empirical investigation of continuous-time equity return models. *J Financ* 57:1239–1284
- Antonakakis N (2012) Exchange return co-movements and volatility spillovers before and after the introduction of Euro. *J Int Financ Mark Inst Money* 22:1091–1109
- Asgharian H, Bengtsson C (2006) Jump spillover in international equity markets. *J Financ Economet* 4:167–203
- Asgharian H, Nossman M (2011) Risk contagion among international stock markets. *J Int Money Financ* 30:22–38
- Bakshi G, Cao C, Chen Z (1997) Empirical performance of alternative option pricing models. *J Financ* 52:2003–2049
- Barndorff-Nielsen O, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. *J Financ Economet* 4:1–30
- Barunik J, Vacha L (2018) Do co-jumps impact correlations in currency markets? *J Financ Mark* 37:97–119
- Barunik J, Kocenda E, Vacha L (2017) Asymmetric volatility connectedness on the forex market. *J Int Money Financ* 77:39–56
- Bates D (2000) Post '87 crash fears in S&P500 future options. *J Economet* 94:181–238
- Bengtsson C (2006) International jumps in returns. Applications of bayesian econometrics to financial economics. *Lund Economic Studies* 64
- Bollerslev T, Law T, Tauchen G (2008) Risk, jumps, and diversification. *J Economet* 144:234–256
- Broadie M, Chernov M, Johannes M (2006) Model specification and risk premia: evidence from future options. *J Financ* 3:1453–1490
- Brooks M, Deans C, Wallis P, Watson B, Wyrzykowski W (2013) Developments in foreign exchange and OTC derivatives markets. *RBA Bulletin*, Sydney
- Chan KF, Powell JG, Treepongkaruna S (2014) Currency jumps and crises: do developed and emerging market currencies jump together? *Pac Basin Financ J* 30:132–157
- Chatrath A, Miao H, Ramchander S, Villupuram S (2014) Currency jumps, cojumps and the role of macro news. *J Int Money Financ* 40:42–62
- Chernov M, Ghysel E (2000) A study towards a unified approach to the joint estimation of objective and risk-neutral measures for the purpose of options valuation. *J Financ Econ* 56:407–458
- Das SR, Uppal R (2004) Systematic risk and international portfolio choice. *J Financ* 59:2809–2834
- Devereux MB, Engel C (2003) Monetary policy in the open economy revisited: price setting and exchange rate flexibility. *Rev Econ Stud* 70:765–783

- Diebold FX, Yilmaz K (2012) Better to give than to receive: predictive directional measurement of volatility spillovers. *Int J Forecast* 28:57–66
- Diebold FX, Yilmaz K (2015) Financial and macroeconomic connectedness: a network approach to measurement and monitoring. Oxford University Press, Oxford
- Duffie D, Pan J, Singleton K (2000) Transform analysis and asset pricing for affine jump diffusions. *Econometrica* 68:1343–1376
- Eraker B (2004) Do equity prices and volatility jump? Reconciling evidence from spot and option prices. *J Financ* 59:1367–1403
- Eraker B, Johannes M, Polson N (2003) The impact of jumps in volatility and returns. *J Financ* 53:1269–1300
- Huang X, Tauchen G (2005) The relative contribution of jumps to total price variance. *J Financ Economet* 3:456–499
- Jawadi F, Louhichi W, Cheffou AI (2015) Testing and modeling jump contagion across international stock markets: a nonparametric intraday approach. *J Financ Mark* 26:64–84
- Johannes M (2004) The statistical and economic role of jumps in interest rates. *J Financ* 59:227–260
- Kitamura Y (2010) Testing for intraday interdependence and volatility spillover among the Euro, the Pound and the Swiss franc markets. *Res Int Bus Financ* 24:158–171
- Lee SS, Mykland PA (2008) Jumps in financial markets: a new nonparametric test and jump dynamics. *Rev Financ Stud* 21:2535–2563
- McMillan DG, Speight AE (2010) Return and volatility spillovers in the three Euro exchange rates. *J Econ Bus* 62:79–93
- Nikkinen J, Sahlstrom P, Vahamaa S (2006) Implied volatility linkages among major European currencies. *J Int Financ Mark Inst Money* 16:87–103
- Pan J (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. *J Financ Econ* 63:3–50
- Patton A, Sheppard K (2015) Good volatility, bad volatility: signed jumps and the persistence of volatility. *Rev Econ Stat* 97:683–697
- Pesaran HH, Shin Y (1998) Generalized impulse response analysis in linear multivariate models. *Econ Lett* 58:17–29
- Piazzesi M (2003) Bond yields and the federal reserve. *J Polit Econ* 113:311–344
- Pukthuanthong K, Roll R (2015) Internationally correlated jumps. *Rev Asset Pric Stud* 5:92–111
- Segal G, Shaliastovich I, Yaron A (2015) Good and bad uncertainty: macroeconomic and financial market implications. *J Financ Econ* 117:369–397
- Zhou C, Wu C, Wang Y (2019) Dynamic portfolio allocation with time-varying jump risk. *J Empir Financ* 50:113–124



# Modeling Currency Exchange Data with Asymmetric Copula Functions



Emel Kızılok Kara, Sibel Açık Kemaloğlu, and Ömer Ozan Evkaya

**Abstract** In the fields of economics and finance, there are data sets with dependent structures that can be modeled symmetrically or asymmetrically. Analyzing the asymmetrically dependent data with a symmetric model can result in inaccurate financial decisions. Besides, the effect of any event such as the financial crisis on international financial returns can be captured more accurately with asymmetric models. Recent studies have revealed that asymmetric dependent structures can be observed in exchange rates. While dependency structures for a financial data set can be modeled with copula functions efficiently, asymmetric dependencies can be modeled with directional copula functions. In the literature, there are some asymmetric copula models constructed in different ways to model directional dependence. The aim of this study is to model asymmetric exchange rate data with directional dependency measures. For this reason, the dependence among the four currencies traded in US Dollars is investigated using Khoudraji type copula functions. Additionally, the proportions of the total variability between foreign exchange returns are examined in detail.

**Keywords** Directional dependence · Asymmetric copula · Khoudraji copula · Currency exchange rate

---

E. Kızılok Kara (✉)

Department of Actuarial Sciences, Faculty of Arts and Sciences, Kırıkkale University, Kırıkkale, Turkey

e-mail: [emel.kizilok@kku.edu.tr](mailto:emel.kizilok@kku.edu.tr)

S. Açık Kemaloğlu

Department of Statistics, Faculty of Sciences, Ankara University, Ankara, Turkey

e-mail: [acik@ankara.edu.tr](mailto:acik@ankara.edu.tr)

Ö. O. Evkaya

Department of Statistical Sciences, Universita Di Padova, Padova, Italy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_4](https://doi.org/10.1007/978-3-030-85254-2_4)

## 1 Introduction

The copula is a probabilistic modeling method, first described by Sklar (1959) and used for modeling the dependence between random variables. In recent years, copulas have gained popularity especially in areas such as finance, economics, engineering, and actuarial science. Copulas allow modeling the dependency structure of the joint random variables independently from their marginal distributions. Besides, they are very useful whenever the assumption of normality does not hold. For these reasons, copula models are really flexible statistical modeling tools to understand the joint behavior of interrelated random variables. For more details about the foundations of copula theory, the interested reader can follow the books of Joe (1997) and Nelsen (2006).

In recent studies, especially in fields such as finance and economy, the dependence of data sets has been examined with copula models. In addition, whether real data sets such as foreign currency or stock exchange data are symmetric as well as dependency, is a remarkable issue in recent studies. Although copula functions are really applied modeling approaches for financial applications, most of the existing families preserve symmetric dependence patterns, i.e., the exchangeability of the variables. For that reason, as a new alternative, asymmetric copulas appear in the literature since classical symmetric copulas are not suitable for all data sets. Mainly, the inspiring work of all recent efforts is the earlier study of Khoudraji (1995), which allows us to investigate directional dependence between the variables. By the 2000s, many researchers studied asymmetric copulas by introducing different construction methods.

Previously, Liebscher (2008) proposed new methods, close to the study of Khoudraji (1995), for the construction of asymmetric multivariate copulas. A new technique, the product of copulas with powered arguments, is introduced by Durante (2009) to construct an asymmetric copula, by following the results of Liebscher (2008). Recently, as an alternative approach, the convex combination of asymmetric copulas is studied by Wu (2014). More theoretically, Siburg et al. (2016) studied the order of asymmetry of the bivariate copula. Later, Mukherjee et al. (2018) have investigated various combinations of the asymmetric copulas over the car rental data set. With the help of asymmetric copula construction, one can study the directional dependence measure. Mainly, this measure is equivalent to the direction of influence between two dependent random variables (Kim and Kim 2014). In one of the earlier studies, Sungur (2005a) mentioned the importance of the copula regression for modeling directional dependence, which allows us to detect the existence of directional dependence. Considering the asymmetric dependence, the variability of one variable with respect to another can be explained by the directional dependence measurements (Mukherjee et al. 2018; Uhm et al. 2012; Jung et al. 2008; Kim and Kim 2014; Wu 2014). Thereafter, several researchers used this main idea to examine the asymmetry of financial data using an asymmetric version of the Farlie–Gumbel–Morgenstern (FGM) copula family (Jung et al. 2008; Uhm et al.

2012). More recently, the previous limitations of the studies for directional dependence were properly addressed, and a multi-step procedure is proposed for optimal parameter estimation by Kim and Kim (2014).

As an applied contribution, the main aim of the study is to investigate the performance of Khoudraji type asymmetric copulas on financial data and discuss the directional dependence coefficients. In this respect, the considered models in this study are directly attached to the study of Khoudraji (1995). Apart from the use of FGM, widely considered Archimedean families such as Clayton, Gumbel, and Frank copulas are incorporated with the independence copula for the asymmetric models. In order to compare the performance of the models over the exchange rates, a set of comprehensive model selection criteria is examined. To the best of our knowledge, such asymmetric copula-based studies are mainly limited to the exchange rate market, and even it is not established yet for the data sets including Turkish currency data. For that purpose, this study focuses on currency data set, retrieved from the Turkish Central Bank Exchange Rate data repository, between July 15, 2015 and July 15, 2018 (3 years period) including serious downward movements in terms of TRY/USD ratio. Furthermore, the directional dependence measure coefficients might be a good explanatory tool to understand the direction of the influences between the exchange rates. In that respect, the findings of the study can serve as a good introductory application including the Turkish exchange rate data set.

The organization of the articles is summarized under four main sections. After the introduction part, the concept of the asymmetric copula model with its main properties, asymmetric tests and the directional dependence measure is defined in Sect. 2. To illustrate the performance of the considered models, Sect. 3 provides main findings for the parameter estimation and the directional dependence measure interpretation over both simulated and considered exchange currency data sets. Finally, the main conclusion of the study with its pros and cons is summarized in Sect. 4.

## 2 The Measurement of the Directional Dependence by the Asymmetric Copula

### 2.1 Construction of Asymmetric Copulas

Sklar's theorem (Sklar 1959) states that the bivariate joint distribution function with marginals  $F$  and  $G$  can be uniquely expressed with a copula  $C : I^2 \rightarrow I$  in the form  $H(x, y) = C(F(x), G(y))$  where  $\forall(x, y) \in IR$  and  $I = [0, 1]$ . Here, the bivariate copula with uniform marginal distribution ( $F(x) = U$  and  $G(y) = V$ ) is written as  $C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v))$  and it satisfies the distribution function properties (Nelsen, 2006).

On the other side, asymmetric copulas can be defined as below, according to the theorem given by Durante (2009). The theorem states that the function  $C_{\alpha, \beta} : I^2 \rightarrow I$  expressed by Eq. (1) is an asymmetric copula for all  $\alpha, \beta \in (0, 1)$  with  $C_1$  and  $C_2$

symmetric families.

$$C_{\alpha,\beta}(u, v) = C_1\left(u^{\bar{\alpha}}, v^{\bar{\beta}}\right)C_2\left(u^\alpha, v^\beta\right) \tag{1}$$

where  $\bar{\alpha} = 1 - \alpha$  and  $\bar{\beta} = 1 - \beta$ . Here, if  $\alpha \neq 1/2$  and  $\beta \neq 1/2$  then  $C_{\alpha,\beta}$  is asymmetric, whereas  $\alpha = \beta$  implies that  $C_{\alpha,\beta}$  is symmetric. Equivalently, based on  $C_1$  and  $C_2$  as both symmetrical copulas, if  $C_{\alpha,\beta}(u, v) = C_{\beta,\alpha}(v, u)$  then  $C_{\alpha,\beta}$  is the symmetric copula. However, if  $C_{\alpha,\beta}(u, v) \neq C_{\beta,\alpha}(v, u)$  then  $C_{\alpha,\beta}$  is called an asymmetric copula. Mukherjee et al. (2018) state that, in asymmetric models, while the correlation values for selected  $(\alpha, \beta)$  parameters in asymmetric models are different from the values calculated for  $(\beta, \alpha)$ , they are the same in symmetric models (that is  $\rho(C_{\alpha,\beta}) = \rho(C_{\beta,\alpha})$  and  $\tau(C_{\alpha,\beta}) = \tau(C_{\beta,\alpha})$ ). Here, Spearman’s  $\rho$  and Kendall’s  $\tau$  correlation coefficients depending on the copula are expressed as  $\rho_C = 12 \int_0^1 \int_0^1 C(u, v) dudv - 3$  and  $\tau_C = 4 \int_0^1 \int_0^1 C(u, v) dudv - 1$ , respectively.

For testing the symmetry of data, Cramer–von Mises statistic was used, defined by Genest et al. (2012) as  $S_n^* = \int_0^1 \left\{ \widehat{C}_n(u, v) - \widehat{C}_n(v, u) \right\}^2 d\widehat{C}_n(v, u)$ . If the p-value is less than 0.05, the data has a symmetrical pattern, otherwise, it represents asymmetric structure. In our study, above-mentioned structure difference has been checked “exchTest” function in the “copula” package (Hofert et al. 2020b).

For alternative models, Khoudraji type asymmetric copula families are used. They are constructed by selecting the independent copula for  $C_1$  and one of the Archimedean copulas (Clayton, Frank, Gumbel) for  $C_2$  in Eq. (1) (Khoudraji 1995, Genest et al. 1998). The considered asymmetric copula models will be indicated with C12, C13, and C14 abbreviations, respectively, later in Sect. 3 for simplicity. In the above construction, independent copula and Archimedean copula families with the dependence parameter  $(\theta)$  are defined as follows (Nelsen 2006):

- (1) Independent:  $C(u, v) = uv$
- (2) Clayton:  $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}, \theta \in (0, \infty)$
- (3) Frank:  $C(u, v) = -\frac{1}{\theta} \log \left\{ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)} \right\}, \theta \in R \setminus \{0\}$
- (4) Gumbel:  $C(u, v) = \exp\{-[(-\log(u))^\theta + (-\log(v))^\theta]^\frac{1}{\theta}\}, \theta \in [1, \infty)$

## 2.2 The Measurement of Directional Dependence

If the asymmetric dependence between the  $(U, V)$  pair is evaluated by copula regression functions, the regression functions for  $U$  and  $V$  will not be the same. This means that the direction of dependence from  $U$  to  $V$  or  $V$  to  $U$  will be different for asymmetric copulas. For this purpose, to model the directional dependence in joint behavior, it is preferred to use the asymmetric copula models. On the other hand,

since directional dependence is not affected by marginals, the resulted joint behavior may differ (See details in Kim and Kim 2014).

By Sungur (2005a, 2005b), the directional dependency coefficients for a copula function  $C$  with uniform marginals over  $[0, 1]$  are expressed in terms of copula regression functions as follows:

$$\begin{aligned} \rho_{U \rightarrow V}^{(2)} &= \frac{\text{Var}(r_{V|U}(U))}{\text{Var}(V)} = 12E[(r_{V|U}(u))^2] - 3, \\ \rho_{V \rightarrow U}^{(2)} &= \frac{\text{Var}(r_{U|V}(V))}{\text{Var}(U)} = 12E[(r_{U|V}(v))^2] - 3 \end{aligned}$$

where  $r_{V|U}(u)$  and  $r_{U|V}(v)$  are copula regression functions. Here,  $\rho_{U \rightarrow V}^{(2)}$  indicates that the total change in  $V$  can be explained by the copula regression of  $V$  on  $U$ . Above copula regression functions are defined as  $r_{V|U}(u) = E[V|U = u] = 1 - \int_0^1 C_u(v)dv$  and  $r_{U|V}(v) = E[U|V = v] = 1 - \int_0^1 C_v(u)du$  where the conditional distribution functions are denoted by  $C_u(v)$  and  $C_v(u)$  having the definitions of  $C_u(v) \equiv P(V \leq v|U = u) = \frac{\partial C(u,v,\phi)}{\partial u}$  and  $C_v(u) \equiv P(U \leq u|V = v) = \frac{\partial C(u,v,\phi)}{\partial v}$  for the parameter set,  $\phi = (\theta, \alpha, \beta)$ .

It may not always be possible to easily find the closed-form of copula regression functions. For this reason, it is calculated approximately with the equations  $\tilde{r}_{V|U}(u) = 1 - \frac{1}{S} \sum_{s=1}^S C_u(v_s)$  and  $\tilde{r}_{U|V}(v) = 1 - \frac{1}{S} \sum_{s=1}^S C_v(u_s)$  over the pseudo-observations,  $(u_s, v_s) \in (0, 1)^2$ . Therefore, the following approximate calculations are used for the directional dependency coefficients:

$$\tilde{\rho}_{U \rightarrow V}^{(2)} = \frac{12}{S} \sum_{s=1}^S (\tilde{r}_{V|U}(u_s))^2 - 3 \text{ and } \tilde{\rho}_{V \rightarrow U}^{(2)} = \frac{12}{S} \sum_{s=1}^S (\tilde{r}_{U|V}(v_s))^2 - 3.$$

### 2.3 Parameter Estimation and Model Selection

In this study, the maximum pseudo-likelihood function (MPL) method studied by Kim and Kim (2014) was used for the parameter estimation. In practice, it has a common usage since it is not affected by the selection of marginals. Firstly, the pseudo-observations  $(u_i, v_i), \{i = 1, \dots, n\} \in (0, 1)^2$  are created in the form  $u_i = R_i/(n + 1)$  and  $v_i = S_i/(n + 1)$ . Here,  $R_i$  and  $S_i$  show the ranks of  $x_i$  and  $y_i$ , respectively. Then, parameter estimation is derived based on Nelder–Mead optimization technique for the parameter set,  $\phi = (\theta, \alpha, \beta)$  using the MPL function defined as

$$l(\phi) = \log \prod_{i=1}^n c(u_i, v_i, \phi) = \sum_{i=1}^n \log c(u_i, v_i, \phi)$$

Here,  $c(u_i, v_i, \phi)$  is the copula density function defined by  $c(u, v, \phi) = \frac{\partial^2 C(u,v,\phi)}{\partial u \partial v}$  for the asymmetric copula,  $C(u_i, v_i, \phi)$ .

Furthermore, the directional dependence criteria mentioned in Kim and Kim (2014) was used to derive the best model among different candidates. According to these main criteria, firstly the asymmetric copula models with the largest/smallest pair of  $\rho_{U \rightarrow V}^{(2)}$  and  $\rho_{V \rightarrow U}^{(2)}$  directional dependence measures are selected. Thereafter, the goodness of fit (GOF) test is performed to see the compatibility of these selected copulas. For this reason, secondly, the model with the smallest AIC value is considered.

### 3 Numerical Findings

#### 3.1 The Examination of the Asymmetric Dependence for Simulated Data

In this subsection, the changes in directional dependence coefficients for asymmetric models are exemplified by simulation. For this purpose, first of all, parameter values for symmetric and asymmetric copulas corresponding to Spearman’s rho correlations are shown in Table 1. Later, dependent data pairs with a sample size of  $N = 1000$  were generated by simulation from the asymmetric copulas for these parameters. The parameter estimations for these produced data pairs are given in Table 2. Also, the asymmetry test results of these generated data pairs are presented in Table 3. In addition to the test results, the symmetric  $p > 0.05$  and asymmetric  $p < 0.05$  structures can be also observed from the scatter (in Fig. 1) and contour (in Fig. 2) plots for  $\rho = 0.6$ . Finally, directional dependence measures in asymmetric models were calculated to show that there are differences in the dependency directions due to asymmetry (in Table 4). Here, it can be concluded that the directional dependence also increases while the Spearman correlation increases in the positive direction. However, it is noteworthy that the rate of increase of the directional dependence measures from  $u$  to  $v$  and  $v$  to  $u$  is not the same.

**Table 1** The parameter values of symmetric and asymmetric copulas for various correlations

$\rho$	Symmetric Archimedean models			Asymmetric Khoudraji models		
	Clayton ( $\theta$ )	Frank ( $\theta$ )	Gumbel ( $\theta$ )	Clayton ( $\theta = 20$ )	Frank ( $\theta = 30$ )	Gumbel ( $\theta = 20$ )
0.2	3.25	1.2	1.16	$(\alpha = 0.8, \beta = 0.6)$		
0.4	0.75	2.6	1.38	$(\alpha = 0.6, \beta = 0.4)$		
0.6	1.5	4.5	1.75	$(\alpha = 0.4, \beta = 0.2)$		
0.8	3.25	8	2.6	$(\alpha = 0.2, \beta = 0.1)$		

**Table 2** Parameter estimation results for asymmetric data pairs generated by simulation ( $N = 1000$ )

Model	$\rho$	$\theta$	$\alpha$	$\beta$
C12	0.2	20.5259	0.7872	0.5887
	0.4	16.5961	0.5873	0.3940
	0.6	21.4162	0.4092	0.2116
	0.8	19.2101	0.2091	0.1091
C13	0.2	33.3189	0.7638	0.6033
	0.4	32.9412	0.6044	0.4021
	0.6	32.7945	0.4074	0.2048
	0.8	32.4242	0.2037	0.1032
C14	0.2	19.0901	0.8126	0.6252
	0.4	20.4289	0.6078	0.4086
	0.6	20.0317	0.4029	0.2033
	0.8	20.2696	0.2086	0.1034

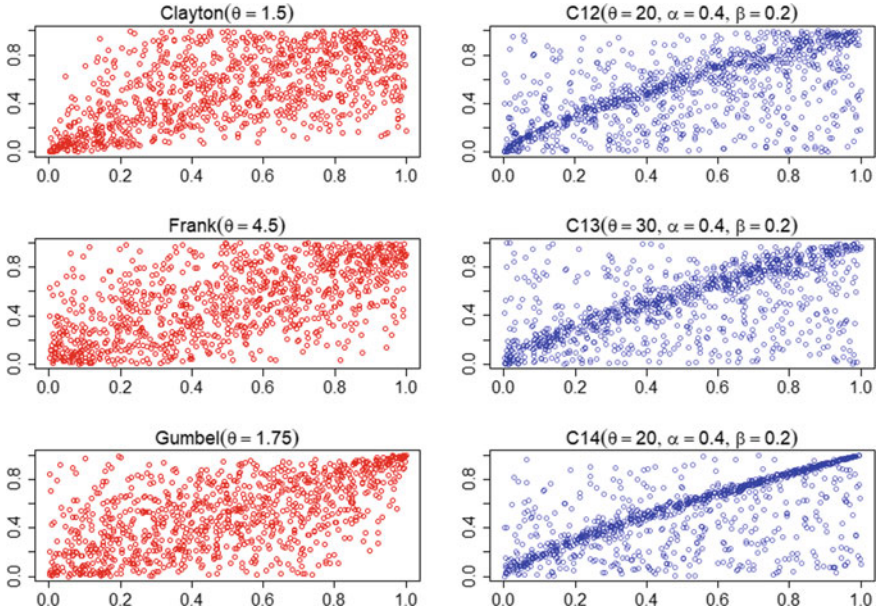
**Table 3** The asymmetry test results for data pairs simulated from symmetric and asymmetric models: Cramer–von Mises statistics and (p-values)

$\rho$	Symmetric Archimedean models			$\rho$	Asymmetric Khoudraji models		
	Clayton	Frank	Gumbel		C12	C13	C14
0.2	0.0194 (0.8077)	0.02595 (0.5899)	0.0147 (0.9665)	0.2	0.10565 (0.003497)	0.088239 (0.005495)	0.13149 (0.001499)
0.4	0.0115 (0.9935)	0.0350 (0.2473)	0.0138 (0.9665)	0.4	0.25793 (0.0004995)	0.2195 (0.0004995)	0.41101 (0.0004995)
0.6	0.0112 (0.9715)	0.0190 (0.6329)	0.0199 (0.5819)	0.6	0.40952 (0.0004995)	0.4109 (0.0004995)	0.66216 (0.0004995)
0.8	0.0192 (0.1733)	0.0198 (0.2443)	0.0096 (0.9685)	0.8	0.050252 (0.007493)	0.10805 (0.0004995)	0.22599 (0.0004995)

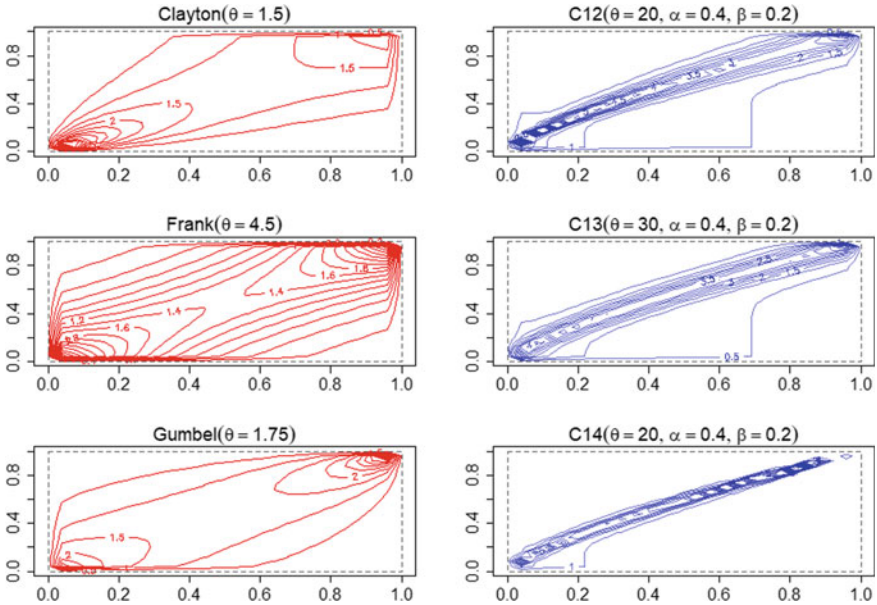
### 3.2 Examining Asymmetric Dependence for Currency Data

The daily foreign exchange market data, used in this study, is obtained from the CBRT Exchange Rate Service between July 15, 2015 and July 15, 2018. Five currencies traded in US Dollars (USD); TRY (TRY/USD), EURO (EUR/USD), RUB (RUB/USD), CNY (CNY/USD) and GBP (GBP/USD) are used and the corresponding four foreign currency pairs formed. (TRY, EUR), (TRY, RUB), (TRY, CNY), (TRY, GBP) are analyzed using the directional copula functions mentioned here to check whether at least one of these four pairs has any directional dependencies.

As prior calculation, the rate of return is obtained by using the logarithmic ratio formula defined as  $r_{i,t} = 100 \times \log(\frac{x_{i,t}}{x_{i,t-1}})$ . To illustrate, the scatter plots for exchange



**Fig. 1** The scatter plots for symmetric (left column) and asymmetric copulas (right column) ( $\rho = 0.6, N = 1000$ )



**Fig. 2** The contour plots for symmetric (left column) and asymmetric copulas (right column) ( $\rho = 0.6, N = 1000$ )

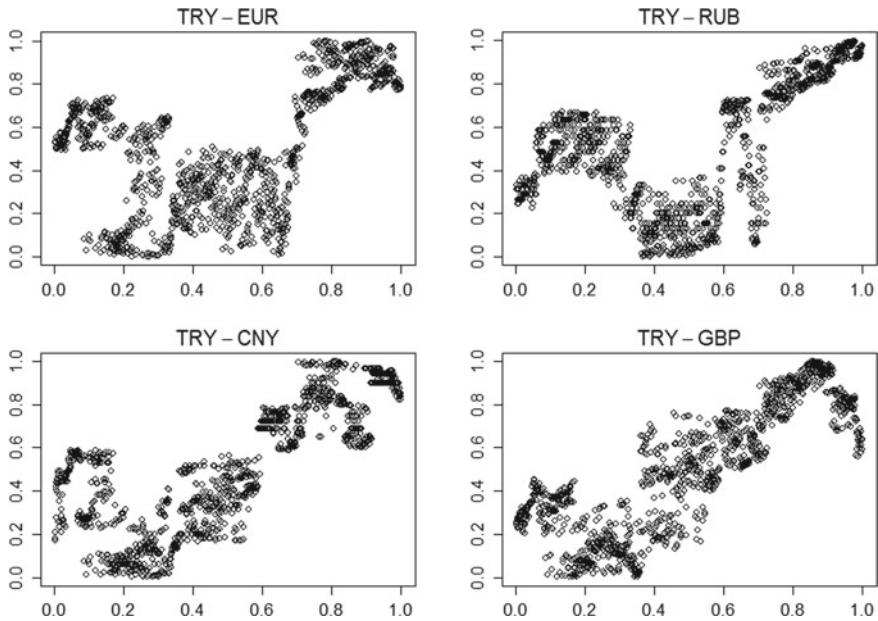


**Table 4** Directional dependence coefficients of asymmetric Khoudraji copula models

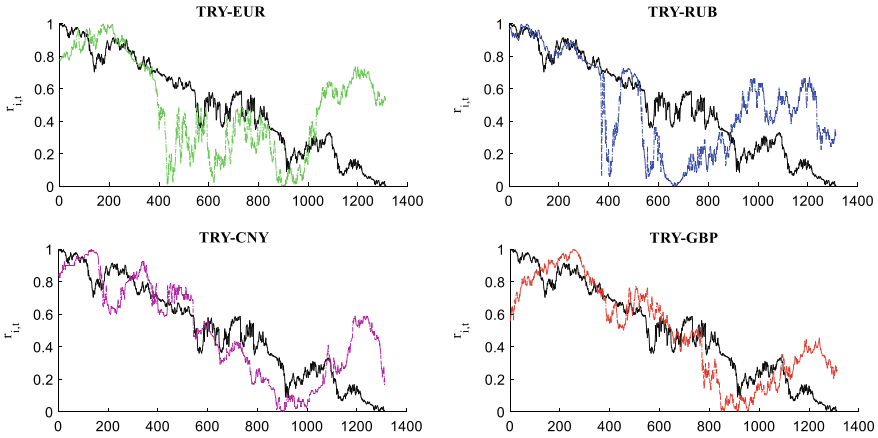
	C12		C13		C14	
$\rho$	$\rho_{u \rightarrow v}^{(2)}$	$\rho_{v \rightarrow u}^{(2)}$	$\rho_{u \rightarrow v}^{(2)}$	$\rho_{v \rightarrow u}^{(2)}$	$\rho_{u \rightarrow v}^{(2)}$	$\rho_{v \rightarrow u}^{(2)}$
0.2	0.0691	0.0688	0.0495	0.0097	0.0504	0.0068
0.4	0.1531	0.3117	0.1568	0.1435	0.2590	0.0801
0.6	0.4087	0.3581	0.3457	0.3512	0.4425	0.2910
0.8	0.4512	0.7198	0.5711	0.6357	0.6350	0.6234

rates according to USD are given in, Fig. 3 and the plots of  $r_{i,t}$  (rate of return) between countries are given in Fig. 4. It is clear from these plots that the direction of the relationship between Turkey’s exchange rates and the other countries’ currency exchange rates is positive. Also, the correlation coefficients for (TRY, GBP), (TRY, CNY), (TRY, RUB), and (TRY, EUR) country pairs were found to be 0.8298, 0.7710, 0.5373, and 0.4930, respectively.

Table 5 provides basic descriptive statistics of  $r_{i,t}$  where it represents not a symmetrical pattern. Here, the skewness and the kurtosis values differ from zero, although the mean and median values for each are almost the same. It is also noticed that GBP and RUB have the largest and smallest rates of return, respectively, against the USD. For the symmetry test, the Cramer–von Mises statistics ( $S_n^*$ ) and p-values described in Sect. 2 are given in Table 6. The data pairs are not symmetric since the



**Fig. 3** The scatter plots for exchange rate pairs



**Fig. 4**  $r_{i,t}$  graphs, x-axis is the time, y-axis is  $r_{i,t}$  (black curve for TRY/USD)

**Table 5** The descriptive statistics for  $r_{i,t}$

	Mean	Median	Min	Max	St.D	Skewness	Kurtosis
TRY	0.3583	0.3429	0.2051	0.5262	0.0852	0.2522	-1.1419
EUR	1.1896	1.1586	1.0387	1.3934	0.1049	0.6163	-1.0135
RUB	0.0197	0.0172	0.0121	0.0316	0.0057	0.9955	-0.6488
CNY	0.1560	0.1571	0.1436	0.1655	0.0063	-0.3377	-1.1654
GBP	1.4583	1.4627	1.2048	1.7166	0.1448	0.0115	-1.2921

**Table 6** Asymmetry test results of exchange rate pairs

	$S_n^*$	p-values
(TRY, EUR)	1.80740	0.0004995
(TRY, RUB)	0.43103	0.0014990
(TRY, CNY)	1.53810	0.0004995
(TRY, GBP)	0.57636	0.0004995

p-value is smaller than 0.05. According to these results, it can be said that there is an asymmetric dependence pattern between the considered foreign exchange returns of pairs.

Goodness of fit test was conducted to compare asymmetric models fitting the data. The results including parameter estimates, log-likelihood (LL), and Akaike information criterion (AIC) values for the asymmetric models (C12, C13, C14) used in the study are presented in Table 7. Besides, for the same models, the corresponding directional dependency coefficients are presented in Table 8.

When the model is selected according to the AIC criterion among the asymmetric models, it is concluded that the best model is C14 for TRY-RUB, and C13 for

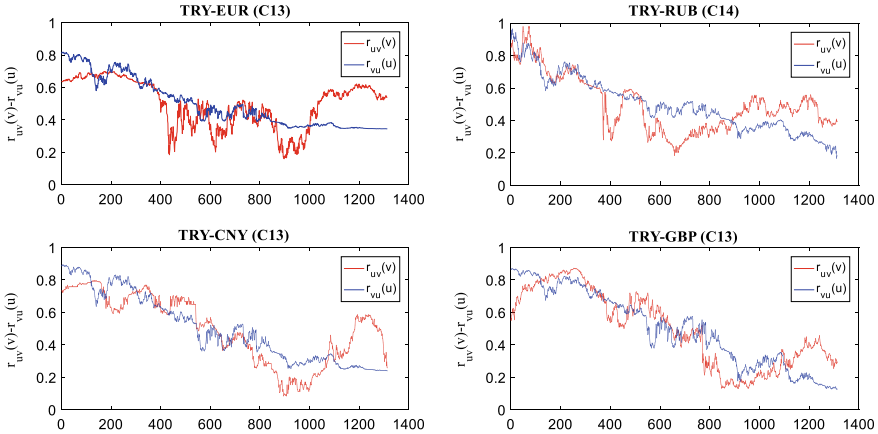
**Table 7** The parameter estimations, LL, and AIC values for asymmetric copula models

	TRY-EUR				TRY-RUB		
Parameters	C12	C13	C14	Parameters	C12	C13	C14
$\hat{\theta}$	15.171	13.578	1.9102	$\hat{\theta}$	10.449	3.5258	1.7281
$\hat{\alpha}$	0.6573	0.8639	0.9989	$\hat{\alpha}$	0.2968	0.9998	0.9523
$\hat{\beta}$	0.4302	0.4756	0.5844	$\hat{\beta}$	0.9799	0.9998	0.9996
LL	273.1	295.4	211.8	LL	287.3	202.3	351.4
AIC	-540.2	-584.8	-417.6	AIC	-568.6	-398.6	-696.8
	TRY-CNY				TRY-GBP		
Parameters	C12	C13	C14	Parameters	C12	C13	C14
$\hat{\theta}$	8.7313	12.404	2.5207	$\hat{\theta}$	7.8396	7.9300	2.1278
$\hat{\alpha}$	0.9138	0.9997	0.9858	$\hat{\alpha}$	0.8584	0.9996	0.8923
$\hat{\beta}$	0.5684	0.7066	0.7356	$\hat{\beta}$	0.5930	0.9995	0.9998
LL	546.8	603.2	442.9	LL	465.2	632.6	432.8
AIC	-1087.6	-1200.4	-879.8	AIC	-924.4	-1259.2	-859.6

**Table 8** Spearman’s  $\rho$  and directional dependence coefficients for asymmetric copula models

TRY-EUR ( $\rho_S = 0.4930, \rho_S^2 = 0.2431$ )				TRY-RUB ( $\rho_S = 0.5373, \rho_S^2 = 0.2887$ )			
	C12	C13	C14		C12	C13	C14
$\rho_C$	0.4238	0.4929	0.4663	$\rho_C$	0.3528	0.5094	0.5733
$\rho_C^2$	0.1796	0.2429	0.2175	$\rho_C^2$	0.1244	0.2595	0.3286
$\rho_{u \rightarrow v}^{(2)}$	0.1901	0.2633	0.2481	$\rho_{u \rightarrow v}^{(2)}$	0.1431	0.2589	0.3394
$\rho_{v \rightarrow u}^{(2)}$	0.1814	0.2504	0.2193	$\rho_{v \rightarrow u}^{(2)}$	0.1601	0.2587	0.3423
TRY-CNY ( $\rho_S = 0.7710, \rho_S^2 = 0.5945$ )				TRY-GBP ( $\rho_S = 0.8298, \rho_S^2 = 0.6886$ )			
	C12	C13	C14		C12	C13	C14
$\rho_C$	0.5864	0.7077	0.6460	$\rho_C$	0.5774	0.7961	0.6676
$\rho_C^2$	0.3439	0.5009	0.4173	$\rho_C^2$	0.3334	0.6338	0.4457
$\rho_{u \rightarrow v}^{(2)}$	0.3576	0.5145	0.4395	$\rho_{u \rightarrow v}^{(2)}$	0.3400	0.6332	0.4511
$\rho_{v \rightarrow u}^{(2)}$	0.3577	0.5068	0.4180	$\rho_{v \rightarrow u}^{(2)}$	0.3431	0.6333	0.4596

TRY-EUR, TRY-CNY, and TRY-GBP. Similarly, when the directional dependence measures are taken into account and the model selection is applied according to the criteria explained in Sect. 2.3, the same models are determined. Here, two asymmetric models are selected from Table 8 that provide the smallest and largest pairs of  $\rho_{u \rightarrow v}^{(2)}$  and  $\rho_{v \rightarrow u}^{(2)}$ . Then, the model with the smallest AIC value among these is determined by Table 7. Therefore, by looking at the directional dependency and AIC criteria



**Fig. 5**  $r_{U|V}(v) - r_{V|U}(u)$  graphs for exchange rate pairs according to the selected models ( $u$ : TRY,  $v$ : EUR, RUB, CNY, GBP)

together, the best asymmetric model is C13 for TRY-EUR, C14 for TRY-RUB, C13 for TRY-CNY, C13 for TRY-GBP.

The obtained asymmetric dependencies between currency pairs can be briefly explained according to the results given in Table 8 as follows: Considering the financial dependencies of TRY with respect to foreign exchange returns with other countries as symmetric, it is seen that the rates of change  $\rho_C^2$  relative to each other after any event are 63.38% (GBP), 50.09% (CNY), 32.86% (RUB) and 24.29% (EUR) from largest to smallest. However, when asymmetric models are used to explain this interdependence, it is seen that these total change rates explained by copula regression are different: the explanation rates for the total change in TRY are 63.32% with GBP, 50.68% with CNY, 34.23% with RUB, and 25.04% with EUR. The total change in TRY is explained with the rates of 63.33% with GBP, 50.68% with CNY, 34.23% with RUB, and 25.04% with EUR. On the other hand, total changes in GBP, CNY, RUB, and EUR are explained by TRY with the rates of 51.45, 34, 33.94, and 26.33%, respectively.

Finally, Fig. 5 shows regression changes according to selected models for the currency pairs. The displayed figures revealed that total change rates are the most in between the foreign currency rates of Turkey, England, and China. In addition, it can be concluded that the exchange rates of countries are affected in different directions by possible changes.

## 4 Conclusion

In this study, the asymmetric copula approach is investigated to understand the change in the directional dependency measures, calculated based on the copula regression

function. This method has been applied to both simulated and currency rate data sets to test its performance. In the first part, simulated dependent data pairs, generated from asymmetric Khoudraji copulas, was analyzed. Here, how the asymmetric dependence changes visually are presented with the scatter and the contour plots. Also, it was observed that there was an increase in the directional dependency coefficients according to increased values of Spearman's rho correlation. Secondly, the method was carried out for a foreign exchange rate of returns for different countries. As a starting point, it has been shown that these row currency data pairs are asymmetrically dependent. Therefore, the best asymmetric copula model among the candidate asymmetric Khoudraji copulas was selected in two steps. The final model was determined using both the directional dependency coefficients based on the copula regression function and AIC values. To sum up, the model selection is done as follows: Two models with the largest or smallest calculated values  $\rho_{u \rightarrow v}^{(2)}$  and  $\rho_{v \rightarrow u}^{(2)}$  were determined in the first step. Thereafter, the one with the smallest AIC value was selected as the best model. Here, it has been demonstrated that the choice of the model according to the directional dependence measures may be different.

As a result, the main conclusions drawn from the study can be summarized as follows:

- For asymmetric data pairs produced using different correlations from the simulation and Khoudraji copulas, there have been significant changes in the directional dependence coefficients.
- For a real data set consisting of asymmetric foreign exchange returns, the model selection was made by taking the directional dependency measures into account is different from the model selection made by known methods. In this sense, it is necessary to check whether the data has an asymmetric dependence pattern or not to consider the directional dependence measures based on the copula regression function.

On the other hand, significant differences were observed in the directional dependency coefficients for real currency data. The main reason for this difference might rely on, the periods of work. More clearly, foreign exchange returns of countries may have been affected differently by important events such as the financial crisis. In this sense, the results of the directional dependency coefficients obtained from the study can help investors on how to make decisions about their foreign exchange investments. However, the asymmetric models considered in order to obtain a more comprehensive information on the market are limited here.

For future studies, it is possible to expand the study by adding other asymmetric models by using different construction methods in the literature. Also, by considering the change point determination approach, comparative inferences can be made with directional dependency measures for different periods. In this new framework, it is possible to implement asymmetric copula models over different time periods, determined by change point analysis. Such improvements are on the list of the authors for possible further contributions.

## References

- Durante F (2009) Construction of non-exchangeable bivariate distribution functions. *Stat Pap* 50(2):383–391
- Genest C, Nešlehová J, Quessy JF (2012) Tests of symmetry for bivariate copulas. *Ann Inst Stat Math* 64(4):811–834
- Joe H (1997) *Multivariate models and dependence concepts*. Chapman and Hall, London
- Jung YS, Kim JM, Kim J (2008) New approach of directional dependence in exchange markets using generalized FGM copula function. *Commun Statistics—Simulation Comput* 37(4):772–788
- Khoudraji A (1995) *Contributions à l'étude des copules et à l'automodélisation des valeurs extrêmes bivariées*. PhD thesis, Université de Laval, Québec
- Kim D, Kim JM (2014) Analysis of directional dependence using asymmetric copula-based regression models. *J Stat Comput Simul* 84(9):1990–2010
- Liebscher E (2008) Construction of asymmetric multivariate copulas. *J Multivar Anal* 99:2234–2250
- Mukherjee S, Lee Y, Kim JM, Jang J, Park JS (2018) Construction of bivariate asymmetric copulas. *Commun Stat Appl Methods* 25(2):217–234
- Nelsen RB (2006) *An Introduction to Copulas*, 2nd edn. Springer, New York
- Siburg KF, Stehling K, Stoimenov PA, Eig GNF (2016) An order of asymmetry in copulas, and implications for risk management. *Insur: Math Econ* 68:241–247
- Sklar A (1959) *Functions de repartition an dimensions at leurs marges*. *Publ Inst Statist Univ Paris* 8:229–231
- Sungur EA (2005) A note on directional dependence in regression setting. *Commun Stat Theory Methods* 34:1957–1965
- Sungur EA (2005) Some observations on copula regression functions. *Commun Stat Theory Methods* 34:1967–1978
- Uhm D, Kim JM, Jung YS (2012) Large asymmetry and directional dependence by using copula modeling to currency exchange rates. *Model Assist Stat Appl* 7:327–340
- Wu S (2014) Construction of asymmetric copulas and its application in two-dimensional reliability modelling. *Eur J Oper Res* 238:476–485

# The Joint Tests of the Parity Conditions: Evidence from a Small Open Economy



Kadir Y. Eryiğit and Veli Durmuşoğlu

**Abstract** This chapter presents the set of international parity conditions which are core financial theories related to exchange rates determination and joint tests of the validity of Uncovered Interest Rate Parity (UIP) and Purchasing Power Parity (PPP), two important international conditions of parity, for Turkey-US and Turkey-Euro Area within the multivariate-cointegration framework of (Johansen et al., *Econometrics Journal* 3:216–249, 2000) study, allowing structural breaks such as the global financial crisis of 2007–2009 and implementation of macroprudential policies after the global financial crisis. The cointegration tests statistics reveal two vectors that cointegrate in both systems containing prices, exchange rates, and interest rates for Turkey-US and Turkey-Euro Area for the 2005:1–2009:4 and 2005:2–2010:1 pairs of breaks, respectively. Additionally, for both systems, each parity condition is rejected when it is formulated jointly, which implies that in a financially open economy, asset and commodity market adjustments might be interrelated. Conversely, when each parity condition is formulated as less restrictive for both systems, it is not rejected. This suggests PPP and UIP with proportionality and symmetry conditions in the first and second vectors, respectively.

**Keywords** Exchange rates · Interest rates · Prices · Cointegration · Structural breaks

---

This chapter was derived from the Project titled “Testing International Parity Conditions for Turkey: A Multivariate Cointegration Analysis in the Presence of Structural Breaks”, supported by the International Post-Doctoral Research Fellowship Programme (2018) of the Scientific and Technological Research Council of Turkey (TUBITAK). Additionally, we pay our special thanks to Alfred H. Haug, University of Otago, New Zealand for supervising.

---

K. Y. Eryiğit (✉) · V. Durmuşoğlu  
Department of Econometrics, Bursa Uludağ University, Bursa, Turkey  
e-mail: [kyeryigit@uludag.edu.tr](mailto:kyeryigit@uludag.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_5](https://doi.org/10.1007/978-3-030-85254-2_5)

## 1 Introduction

Topics such as Purchasing Power Parity (PPP), Uncovered Interest Rate Parity (UIP), Interest Rate Parity (IP), and International Fisher Effect (IFE) are heavily disputed in international macroeconomics. PPP indicates that price differentials can have an effect on exchange rates, contributing to rebalancing of the international commodity market in an open economy. UIP, however, speculates that, whilst considering international asset markets, the exchange rates are correlated with the interest rate differentials. Finally, IFE suggests that the price differentials between countries are aligned to the real interest rate differentials. It has been observed that disequilibrium in one market has similar effects in other markets. From this, it may be observed that international parity conditions are interdependent on one another. From the aforementioned statement, it may be concluded that it is an essential method of testing international parity conditions in tandem with each other, to consider international commodity and asset markets together.

International free market conditions affect countries in a way that leads to a balance among a number of economic and financial variables. In situations where the functioning of the markets is not disrupted, equilibrium relationships are generated among the rates of inflation, exchange, and interest. These relationships, known as parity conditions, are at the center of international financial transactions.

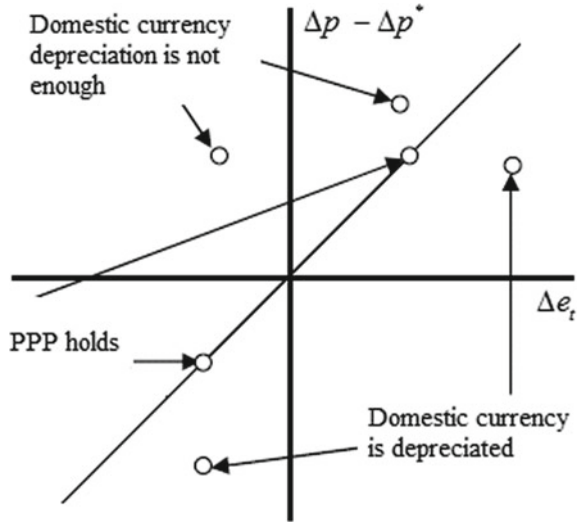
Parity conditions arise as a result of arbitrage activities in international finance, listed as PPP, IP, and IFE. The most widely used one among these conditions is the parity condition (PPP), which is a very controversial issue. PPP (in this case Relative PPP, which refers to the difference between expected change in the exchange rate and inflation rates of the trading partner countries). In short, when the domestic inflation rate is high compared to the foreign inflation, domestic currency's value will decrease and this depreciation will cause an inflationary difference between the two countries as per the following formula:

$$\Delta e = \Delta p - \Delta p^*. \quad (1)$$

Here,  $\Delta p$  stands for the domestic and  $\Delta p^*$  for the foreign country inflation;  $\Delta e$  stands for change in the rate of nominal exchange. For the validity of PPP, (i) goods and services produced by countries must be homogeneous, (ii) international goods markets must be fully competitive, (iii) trade transaction costs must be low or negligible, and (iv) trade barriers between countries must not be presented. The PPP relationship is graphically presented in Fig. 1. Points above the line passing at the angle of  $45^\circ$  from the origin show that the changes in the rate of nominal exchange are the same as the two countries' inflation difference, thus, PPP holds. At the points on the  $45^\circ$  line's right-hand side, domestic currency's value decreases excessively, i.e., the nominal exchange rate's increase is higher compared to the two countries' inflation difference. Points on the  $45^\circ$  line of the upper left side indicate that loss of value of the domestic currency is not sufficient to reveal the two countries' inflation difference.



Fig. 1 PPP relationship



Although PPP appears to be an important tool in explaining long-run exchange rate behavior, it is not enough in the short run. At this instance, PPP’s persistent behavior pattern may be explained as that the general level of prices in the goods market varies less than the nominal exchange rates. In an environment where monetary factors are more dominant than real factors, namely in a period of rapid inflation, PPP relationship’s validity is more evident, but this relationship may not hold in an environment with lower international inflation differentials.

The increased use of flexible exchange rate regimes around the world in the first half of the 1970s resulted in an increase in the concentration of foreign capital movements. This situation has brought up another relationship, known as the interest rate parity relationship. This arises due to the increasing integration in the international capital markets, as the arbitrage activities between the short-run international financial fund markets and the international money markets increased, causing capital to become more liquid. Since then, the interest rate parity theory has become an important tool in explaining the formation of exchange rates.

There is a connection between IP and the expected changes in exchange rates between two countries at a certain time or in a certain region, and also the difference in interest rates. Expected exchange rate changes may be secured for a future time depending on a specific contract. IP condition called Covered Interest Rate Parity, or CIP, reveals a correlation between the forward and spot exchange rates and interest rate differences between two countries. The UIP linked to IFE is derived if the anticipated adjustments in the exchange rate are not secured in compliance with a contract.

CIP states that the differential return from the forward interest rate between two countries should be equal. Otherwise, there will be quite a lot of arbitrage opportunities. In order to eliminate these arbitrage opportunities, the following condition

between interest rates and the forward and spot exchange rates is required to be valid in both countries:

$$f_{0,1} - e_0 = i - i^* \tag{2}$$

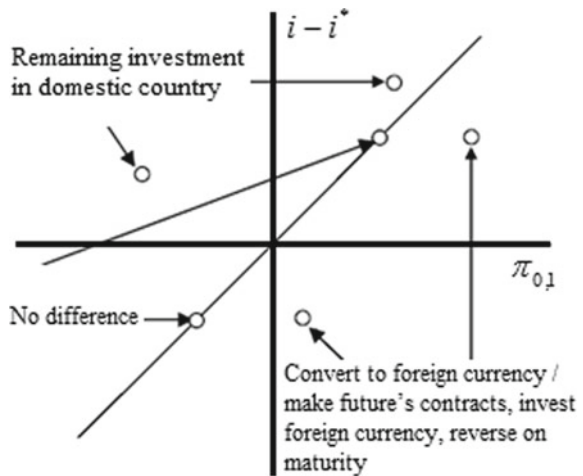
We note that  $f_{0,1}$  represents the forward rate of the currency, delivered in the next period by accounting for that the contract is concluded in the current period. Meanwhile,  $e_0$  represents spot exchange rate,  $i$  represents domestic interest rate, and  $i^*$  represents international interest rate. Premium, or “forward rate”, is given by  $\pi_{0,1} = f_{0,1} - e_0$ , and by equating it to (2), we derive

$$\pi_{0,1} = i - i^* \tag{3}$$

Equations (2) and (3) pinpoint that, to eliminate situations of arbitrage con, the term premium must be directly proportional to the differences in interest rates in both countries.

The CIP relationship is presented graphically in Fig. 2. The points on the line passing through the origin at a 45° angle indicate the points where the matched IP holds and with that, there are neither any arbitrage conditions available, nor there is a difference between investing at home or abroad. Forward premium is small in size compared to domestic and foreign interest rate differences, at points above the same line. Therefore, trading with a forward premium becomes more advantageous compared to interest rate differentials. Difference between domestic and foreign interest rates is smaller than forward premium at points on the right-hand side of the line. At the beginning of the period, it is profitable to convert the currency from local to foreign at current spot exchange rates and make outside investments. Likewise, investors earn interest on the interest rate ( $i^*$ ) by taking advantage of the arbitrage

Fig. 2 CIP relationship



opportunity in the forward contract after converting the foreign return to domestic currency.

On the other hand, as stated by UIP, anticipated changes in the spot exchange rate's interest parity should be the same value as domestic and foreign interest rates. Meanwhile, as mentioned below, there are not any arbitrage opportunities:

$$E_0^e(\Delta e_1) = i - i^* \tag{4}$$

From the equation above,  $E_0^e$ ,  $t = 0$  indicates the mathematical expectations,  $\Delta e_1$  stands for the spot exchange rate, and  $t = 1$  represents the periodic change. If future expectations regarding the spot exchange rates are correct, UIP does not hold, and profitable arbitrage opportunities may be in question. However, if the future expectations regarding the spot exchange rate are not correct, with UIP not being valid, there will be no profit or even loss. Figure 3 presents the UIP relationship graphically. It means that the 45° angle with the correct UIP passing through the origin is valid. As a result, there is no difference between investing in a domestic or a foreign country at any point along this line. Domestic currency's loss of value is bigger than the domestic and foreign interest rate differences at any point right below the line. Converting national currency at latest spot exchange rates and investing in a foreign country to earn interest is profitable if the initial projections are on point. Firms do foreign investments at the end of periods by exchanging currency to domestic, meaning, they get similar results as CIP, provided that the expectations are correct. Domestic currency's anticipated loss of value is small compared to the difference in the interest rates between two countries at any point in the upper left side of the line. After all, if the expectations are on point, the benefit of taking advantage

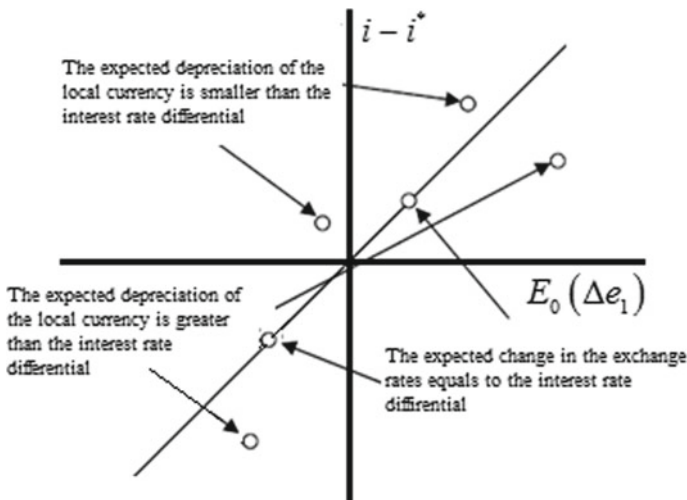


Fig. 3 UIP relationship

of anticipated changes in spot exchange rates will be greater than advantages derived from the interest rate differential.

IFE, obtained as an expression of UIP for different situations, refers to the case that the anticipated changes in exchange rates are the same as interest rate differences between two countries and is expressed as follows:

$$i - i^* = E_0^e(\Delta e_1). \tag{5}$$

The graphical representation of IFE is similar to the graphical representation of UIP as demonstrated in Fig. 3.

Deviations in interest parities may result from (i) transaction costs, (ii) cost of information gathering and process, (iii) government interventions and regulations, (iv) tax differences, (v) financial constraints, and (vi) difficulty in comparisons between assets.

A visual summary of the relationships between international parity conditions is given in Fig. 4.

An approach, which takes into account the relationships between interest rates, exchange rates, and prices, was proposed by the researchers Johansen and Juselius (1992), Juselius (1995), and Juselius and MacDonald (2000,2004), which combined international parity conditions; it is discussed in the next section.

Alternatively, research in international macroeconomics is heavily focused on gaining popularity of emerging market economies in global economic affairs. Turkey has recently attracted international economic attention, where, in addition to the recent political unrest, the country has shown very high growth rates. This chapter

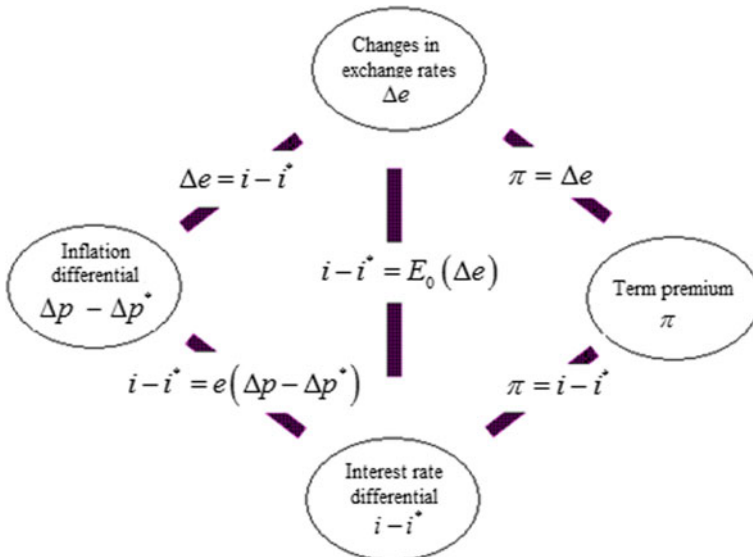


Fig. 4 Relationships in international parity conditions

discusses PPP, UIP, and IFE (international parity conditions) for Turkey-USA and Turkey-Euro Region at the same time. Previous studies which have tested the parity conditions simultaneously for Turkey have been very restricted in terms of their scope. Such an initiative is extremely crucial since some of these interconnected conditions are essential for each other.

Metin (1994), Telatar and Kazdağlı (1998), Sarno (2000), and Erlat (2003) can be listed among the recent significant studies for Turkey. Metin (1994) jointly analyzed UIP and PPP and determined that the yearly data in Turkey on the fixed and flexible exchange rates were not supported by both parties. Based on traditional unit root tests, Telatar and Kazdağlı (1998) provided proof against PPP. In his 2000 study, Sarno used a non-linear method to modeling, strongly advocating the validity of PPP. Erlat's results (2003) indicated that, in the case of the use of fractional-integration and endogenous point estimation approaches, the PPP hypothesis cannot be rejected. In their study Özmen and Gökcan (2004) analyzed the interrelationships between Purchasing Power Parity and Uncovered Interest Rate Parity in Turkey. They used Johansen's cointegration analysis and found that each parity condition was rejected when formulated separately according to the Capital-enhanced Equilibrium Exchange Rate, or CHEER in short. Earlier studies were somewhat restricted and did not acknowledge the presence of structural breaks in long-run relationships.

This study uses quarterly data for the period of 2002:1–2016:4. This period is especially important because since the 2001 financial crisis, Turkey enacted many significant economic changes, including floating and managed float exchange rate regimes, independence of the central bank, banking system adjustments, inflation targeting, and macro prudential policies, which may be seen as structural breaks in the long-run relationship between variables.

Organization of the rest of the chapter is as follows: Following section provides an outline for PPP and UIP. This is followed by a discussion on generic identification of PPP and UIP. The third section introduces the econometric methodology we used. Section 4 provides an ocular analysis for the data and empirical findings. The final section concludes the chapter and presents some comments.

## 2 An Outline of PPP and UIP

Johansen and Juselius (1992) analyzed the long-run foreign transmissions effects between United Kingdom and the rest of the world, whereas Juselius (1995) and Juselius and MacDonald (2000, 2004) considered similar issues for Denmark and Germany, Japan and the US, and the US and Germany. All these studies focused on the long-run UIP and PPP relationships.

PPP is built upon an attractive idea, which suggest that the same traded goods produced in different countries should have an identical price when converted to units of a common currency. If this is not the case, there is room for profitable arbitrage in that good. However, if the exchange rate is free to adjust and there are no impediments to trade, exchange rate's fluctuations should remove the potential for profitable

arbitrage so that PPP condition holds. This adjustment might take time in reality for PPP to hold in long run. Letting  $p_t$  and  $p_t^*$  to represent the logarithms of foreign and national prices,  $e_t$  to represent the nominal exchange rate, calculated as outside currency unit's domestic price, then, at its simplest, PPP implies the following:

$$p_t - p_t^* - e_t = \xi_{1t} \quad (6)$$

where  $\xi_t$  is a stationary, zero mean, and random variable that allows the potential of short-run variations from the fundamental PPP equilibrium described by  $\xi_{1t} = 0$ .

When PPP concentrates on the goods market, we may also expect a capital market condition known as UIP to hold. Suppose that, in addition to the domestically produced good, each country also has a financial asset which, like the good, we assume to be the same apart from the country of origin with domestic and foreign interest rates denoted as  $i_t$  and  $i_t^*$ , respectively. Arbitrage on the capital market creates pressure on the exchange rate, ensuring that an interest rate differential is only associated with a difference between the current exchange rate and the rate projected at  $t$  (time) for  $t + 1$  period. In summary:

$$i_t - i_t^* = \xi_{2t} \quad (7)$$

where  $\xi_{2t} = E_t\{e_{t+1}\} - e_t$  is a stationary, random variable, and  $E_t\{e_{t+1}\}$  is the projected value at  $t$  (time) of the exchange rate in  $t + 1$  (period).

## 2.1 Generic Identification of Parity Conditions

If PPP and UIP are present, we then hope to see two cointegrating relationships among the following five variables:  $p_t$ ,  $p_t^*$ ,  $e_t$ ,  $i_t$  and  $i_t^*$ ; corresponding to PPP and UIP, respectively, where the following should be the cointegrating vectors:  $(1, -1, -1, 0, 0)$  and  $(0, 0, 0, 1, -1)$ . Although this evaluation may be systematically addressed under rank conditions, it should be widely known that the PPP relationship does not include interest rates, and the UIP relationship does not include PPP variables, which is why the PPP and UIP relationships are distinguishable from each other. The requisite condition for identification of  $r = 2$  is that there should be a  $2 - 1 = 1$  restriction on each vector, while there are 4 restrictions on the first vector consisting of 2 exclusions and 2 restrictions on equality and 4 restrictions on the second vector consisting of 3 exclusions and 1 restriction on equality, and each is thus over-identified. The first cointegrating vector: PPP, other coefficients equal to zero; and the second cointegrating vector: UIP, other coefficients equal to zero. The suitably amended  $H_1$  and  $H_2$  matrices are

$$\beta_1 = H_1\varphi_{11} = \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \beta_{31} \\ \beta_{41} \\ \beta_{51} \end{pmatrix} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \varphi_{11} \quad (8)$$

and

$$\beta_2 = H_2\varphi_{12} = \begin{pmatrix} \beta_{12} \\ \beta_{22} \\ \beta_{32} \\ \beta_{42} \\ \beta_{52} \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \varphi_{12}. \quad (9)$$

What is less restrictive but still over-identifying is that the first vector contains the PPP relationship, but the other coefficients are unrestricted, and the second vector contains the UIP relationship, but the other coefficients are unrestricted. In this case, the suitably amended  $H_3$  and  $H_4$  matrices are

$$\beta_1 = H_3\varphi_{11} = \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \beta_{31} \\ \beta_{41} \\ \beta_{51} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \\ \varphi_{31} \end{pmatrix} \quad (10)$$

and

$$\beta_2 = H_4\varphi_{12} = \begin{pmatrix} \beta_{12} \\ \beta_{22} \\ \beta_{32} \\ \beta_{42} \\ \beta_{52} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{pmatrix} \varphi_{12} \\ \varphi_{22} \\ \varphi_{32} \\ \varphi_{42} \end{pmatrix}. \quad (11)$$

Note that the rank conditions are met with these revised restrictions. Having established that the PPP and UIP relationships in different forms are generically identified, the next stage involves seeing if there is empirical confirmation of these structural relationships in the data. The first step of the next stage involves determining the number of vectors that cointegrate, consistent with the data.

### 3 Econometric Methodology

The relationship between PPP and UIP is estimated by using Johansen et al.'s (2000) method, employed to determine whether there are cointegrating relationships among variables. Before proceeding with this test, it is indispensable determining the variables' integrating properties. Johansen et al.'s (2000) method is appropriate since it considers the structural breaks in the model. In this regard, Lee and Strazicich's structural break test (2003) is used to study variables' stationarity properties.

Given the non-stationarity of the variables, the Johansen method (1988) can be used in order to investigate the possibility of long-run relationships among them. For non-stationary variables, however, this method might not be appropriate since it does not take into consideration the structural breaks. Consequently, Johansen et al.'s approach (2000) with slight changes in vector error correction models (VECM) is used. This is demonstrated as shown below:

Let  $Y_t' = [p_t \ p_t^* \ e_t \ i_t \ i_t^*]$  be considered as the I(1) endogenous variables vector with  $r$  relationships that cointegrate. Johansen et al.'s VECM proposition are as follows (2000):

$$\Delta Y_t = \alpha \begin{bmatrix} \beta \\ \gamma \end{bmatrix}' \begin{bmatrix} Y_{t-1} \\ tE_t \end{bmatrix} + \mu E_t + \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \sum_{i=1}^k \sum_{j=2}^q \Psi_{j,i} D_{j,t-i} + \sum_{m=1}^d \Phi_m W_{m,t} + \varepsilon_t \quad (12)$$

Here,  $\Delta$  represents the difference operator,  $k$  represents the length of the lag,  $E_t = [E_{1t} \ E_{2t} \ \dots \ E_{qt}]'$  is a vector of dummy variables  $q$  with  $E_{j,t} = 1$  for  $T_{j-1} + k \leq t \leq T_j$  ( $j = 1, \dots, q$ ) and 0 otherwise; The first  $k$  observation of  $E_{j,t}$  is set to zero and  $E_{j,t}$  is the  $j$ th period's effective sample. Also, the dummy variable for the  $i$ th observation in the  $j$ th period is represented by  $D_{j,t-i}$ , as in  $D_{j,t-i} = 1$  if  $t = T_{j-1} + i$  ( $j = 2, \dots, q, \ t = \dots, -1, 0, 1, \dots$ ) and 0 otherwise. In line with Hendry and Mizon's (1993) method, the mediating dummies  $W_{m,t}$  ( $m = 1, \dots, d$ ) are included to make sure that residuals are fitted well. The cointegrating vector  $\beta$  represents a long run relationship,  $\alpha$  is a vector that represents the speeds of adjustment toward the long-run equilibrium, and  $\gamma = [\gamma_1 \ \gamma_2 \ \dots \ \gamma_q]$  is a matrix of  $(p \times q)$  dimensional parameters of long-run trends; short-run parameters are as follows:  $\mu$  of order  $(p \times q)$ ,  $\Gamma_i$  of order  $(p \times p)$  for  $i = 1, \dots, k$ ,  $\Psi_{j,i}$  of order  $(q \times 1)$  for  $j = 2, \dots, q$  and  $i = 1, \dots, k$ , and  $\Phi_m$  of order  $(q \times 1)$  for  $m = 1, \dots, d$ . The innovations  $\varepsilon_t$  are assumed to be distributed independently and identically with means equal to 0. Moreover, it is symmetrical with the positive variance-covariance matrix  $\Omega$ —that is,  $\varepsilon_t \sim iid(0, \Omega)$ .

Equation (12) indicates the model for linear trend. Here, trends and levels of cointegrating relationships that fluctuate in different periods are represented as  $H_l(r)$ . The likelihood ratio analysis against an  $H_l(p)$  alternative  $r$  cointegration relationship  $H_l(r)$  hypothesis is



$$LR\{H_l(r)|H_l(p)\} = -T \sum_{i=r+1}^p \ln(1 - \hat{\lambda}_i) \quad (13)$$

Here,  $\hat{\lambda}_i$  represents the squared-sample canonical correlation, where  $1 \geq \hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \geq 0$ .

Cointegration relationships do not include a linear trend, yet when there is a breaking level, the model presented in Eq. (12) can be changed as per suggested in the Johansen et al. method (2000), as  $H_c(r)$ . Johansen et al.'s (2000) method reveal critical values for both  $H_l(r)$  and  $H_c(r)$  models using  $\Gamma$ -distribution.

The Likelihood Ratio, or LR, test is used to investigate additional restrictions on the VECM, based on the cointegration rank. These are also implemented by Harris and Sollis (2003) within a standard setting. We expand the LR tests in this analysis, which is in line with what Johansen et al. (2000) and Dawson and Sanjuan (2005) suggested.

## 4 Empirical Findings

Following variables were defined for Turkey-US and Turkey-Euro Area using quarterly data or the period 2002:1–2016:4:  $p_t^{TR}$ , Turkey consumer price index;  $p_t^{US}$ , the United States consumer price index;  $p_t^{EU}$ , Euro Area consumer price index;  $e_t^{\$}$ , Turkish Lira/US Dollar;  $e_t^{\text{€}}$ , Turkish Lira/Euro nominal exchange rates;  $i_t^{TR}$ , Turkey long-term discounted auction rates;  $i_t^{US}$ , the United States long-term treasury bill rates and  $i_t^{EU}$ , Euro Area long-term treasury bill rates. The logs of the variables were used.

As for the structural breaks, variables' (non)stationarity characteristics were analyzed prior to the analysis of their long-run relationships.

Table 1 above presents the Lee and Strazicich's unit root test results with structural breaks. The non-stationary variables presented contain structural breaks. While studying these structural breaks, it may be observed that they relate in general to the 2007–2009 international financial crisis and introduction of macroprudential policies in Turkey around this period. The findings equally suggest that there are two cointegrating relationships in the international parity conditions system to be identified for Turkey-US and Turkey-Euro Area.

Figure 5, as well, displays the graphs with various break points in the model. Inter-break periods are covered in lighter shade in the graphs.

Upon testing the variables' properties of stationarity and break points in the series, we went further to test the cointegrating relationships between the variables to determine the model. Tables 2 and 3 present the trace statistics with significant pairs of breaks 2005:1–2009:4 for Turkey-US and 2005:2–2010:1 for Turkey-Euro Area.<sup>1</sup> From this result, we note that in both cases, the structural breaks in level and trend,

<sup>1</sup> Other structural breaks were omitted from the study because they were determined to be negligible.

**Table 1** Unit root test results

Series	Model	Lag	Break times	$\lambda$	$t$ -statistics	Critical values (5%)
$p_t^{TR}$	Model C	4	2005:1	0.2	-4.61	-5.74
			2009:4	0.6		
$p_t^{US}$	Model C	1	2008:1	0.4	-4.69	-5.67
			2011:3	0.6		
$p_t^{EU}$	Model C	1	2008:2	0.4	-5.37	-5.67
			2011:2	0.6		
$e_t^S$	Model C	3	2007:2	0.4	-5.43	-5.65
			2014:3	0.8		
$e_t^E$	Model C	3	2005:2	0.2	-4.59	-5.74
			2010:1	0.6		
$i_t^{TR}$	Model C	2	2005:4	0.2	-5.51	-5.74
			2009:2	0.6		
$i_t^{US}$	Model C	1	2011:1	0.6	-4.55	-5.73
			2013:2	0.8		
$i_t^{EU}$	Model C	3	2007:1	0.4	-4.88	-5.65
			2014:2	0.8		

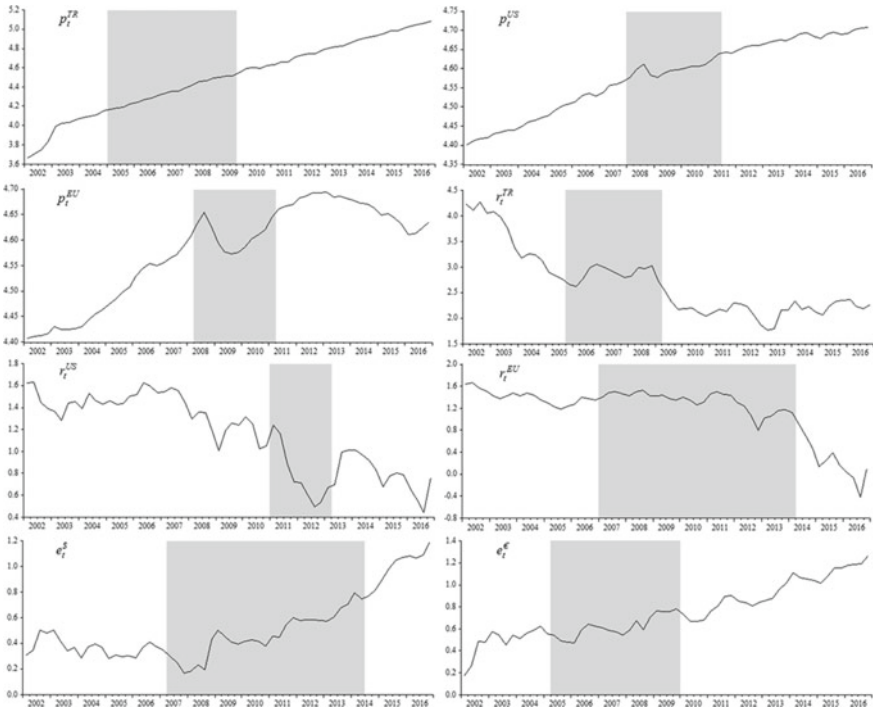
Note Critical values at 5% significance level were retrieved from Lee and Strazicich (2003)

$H_l(r)$  are broken, indicating that breaks in levels and trends are favored in calculating the long-run relationships in the system, instead of  $H_c(r)$ . We also equally determined the lag length for two different structural breaks, shown as  $k = 2$ .<sup>2</sup>

As seen in Tables 2 and 3, we note that the structures of the break pairs for Turkey-US and Turkey-Euro were 2005:1–2009:4 and 2005:2–2010:1, respectively, with two cointegrating vectors at the significance level of 5%. Considering that the residuals are normally distributed in this model, there is no need to include the intervention dummies.<sup>3</sup> Tables 4 and 5, on the other hand, display the LR statistics of the VECM, and from Tables 3 and 4, each variable is maintained in the cointegration space due to the individual exclusion test statistics. This indicates that combining variables with broken levels and trends make the models stationary. Moreover, domestic prices and nominal exchange rates were determined as endogenous variables for both models according to the weak exogeneity test statistics, whereas the others were determined as weak exogenous. Since there is not any difference among sub-samples, the test statistics with existing long-run structural breaks in equilibrium situations often cause the null hypothesis to be rejected. Consequently, for both models, the break pairs of

<sup>2</sup> To determine the optimal lag length, the Akaike Information Criterion's, or AIC's, minimum values were used.

<sup>3</sup> The test results for Multivariate skewness, kurtosis, and joint are: 0.429( $p = 0.807$ ), 6.989( $p = 0.031$ ), and 7.418( $p = 0.155$ ) for Turkey-US model, and 1.821 ( $p = 0.402$ ); 1.211 ( $p = 0.545$ ), and 3.032 ( $p = 0.552$ ), respectively. The test statistics suggest normally distributed models.



**Fig. 5** Time series graphs with structural breaks

**Table 2** Trace statistics for Turkey-US with pair of breaks 2005:1–2009:2

Pair of breaks	$H_0 (H_1)$	Model $H_1(r)$
2005:1–2009:4	$r = 0 (r \geq 1)$	159.26 (137.71)
	$r = 1 (r \geq 2)$	106.91 (103.75)
	$r = 2 (r \geq 3)$	68.52 (73.82)
	$r = 3 (r \geq 4)$	40.18 (47.63)
	$r = 4 (r \geq 5)$	13.58 (24.69)

As per the description by Johansen et al. (2000),  $\Gamma$ -distribution was used to approximate the critical values presented in parentheses at 5% significance level

2005:1–2009:4 and 2005:2–2010:1 are statistically significant with some effects on the long-run relationship.

Tables 4 and 5 also demonstrated the LR tests of the restrictions of the  $\beta_1 = H_1\varphi_{11}$  and  $\beta_2 = H_2\varphi_{12}$  hypotheses for Turkey-US and Turkey-Euro Area models as in Eqs. (8) and (9). For both models, these hypotheses are strongly rejected with the LR statistics of 28.438 and 27.094 (with  $p = 0.000$  and  $p = 0.000$  respectively) distributed as  $\chi_{(6)}^2$ , respectively. Thus, parity conditions are rejected when jointly

**Table 3** Trace statistics for Turkey-Euro Area with pair of breaks 2005:2–2010:1

Pair of breaks	$H_0(H_1)$	Model $H_i(r)$
2005:2–2010:1	$r = 0 (r \geq 1)$	155.81 (138.27)
	$r = 1 (r \geq 2)$	105.28 (104.27)
	$r = 2 (r \geq 3)$	64.25 (74.25)
	$r = 3 (r \geq 4)$	37.32 (47.95)
	$r = 4 (r \geq 5)$	17.16 (24.85)

As per the description by Johansen et al. (2000),  $\Gamma$ -distribution was used to approximate the critical values presented in parentheses at 5% significance level

formulated. This suggests an adjustment in the commodity and asset markets and a possible interdependency.

Meanwhile, all parity condition formulations are less restrictive in both models as in Eqs. (10) and (11). The LR tests of the restrictions of  $\beta_1 = H_3\varphi_{11}$  and  $\beta_2 = H_4\varphi_{12}$  are not rejected with the LR statistics of 3.567 and 1.219 (with  $p$ -values of 0.058 and 0.269, respectively) distributed as  $\chi^2_{(1)}$ , respectively. These results support the PPP and UIP conditions in the first and second vectors with proportionality and symmetry.

## 5 Conclusion and Comments

This chapter employs Johansen et al.'s method of multivariate cointegration from 2000 to examine the PPP and UIP for Turkey-US and Turkey-Euro Area. Such a framework permits the occurrences of structural breaks within the data, like the global financial meltdown of 2007–2009. It also allows the implementation of macroprudential policies after this global financial crisis. Particularly in Turkey, too much flow of capital since 2010 has caused the current account balance to deteriorate, thereby allowing the Turkish Lira to appreciate and push up credit. Meanwhile, the two conditions of PPP and UIP signify an equilibrium condition in the capital and international commodity markets. This is because one market's disequilibrium can have a spill-over effect on the other. In this regard, it would not be an accurate description to state that one parity condition is the level of equilibrium for an entire market. Therefore, it is important to test these conditions in the international financial market.

From the cointegration result, two cointegrating vectors were obtained in both systems containing prices, exchange rates, and interest rates in both Turkey-US and Turkey-Euro Area. For the fact that Turkey is considered as a small open economy among commodity and capital markets, a weak exogeneities in all variables in both systems is not surprising. Specifically, the endogeneities in domestic prices and exchange rates were linked to the fact that they passed-through other policies while causing a disinflation situation.

**Table 4** VECM restrictions test results and identified long-run equations with adjustment coefficients for Turkey-US

Individual exclusion of:	$H_0$	$L$ -statistics	Weak exogeneity of:	$H_0$	$L$ -statistics	Structural breaks	$H_0$	$L$ -statistics				
$p_t^{TR}$	$\beta_{p^{TR}} = 0$	6.211 (0.045)	$p_t^{TR}$	$\alpha_{p^{TR}} = 0$	10.803 (0.000)	2005:1	$\gamma_1 = \gamma_2$	32.608 (0.000)				
$p_t^{US}$	$\beta_{p^{US}} = 0$	19.025 (0.000)	$p_t^{US}$	$\alpha_{p^{US}} = 0$	2.391 (0.303)							
$e_t^s$	$\beta_{e^s} = 0$	16.129 (0.000)	$e_t^s$	$\alpha_{e^s} = 0$	6.795 (0.033)							
$i_t^{TR}$	$\beta_{i^{TR}} = 0$	7.854 (0.019)	$i_t^{TR}$	$\alpha_{i^{TR}} = 0$	5.597 (0.061)	2009:4	$\gamma_2 = \gamma_3$	32.011 (0.000)				
$i_t^{US}$	$\beta_{i^{US}} = 0$	6.851 (0.032)	$i_t^{US}$	$\alpha_{i^{US}} = 0$	0.314 (0.854)							
Standardized eigenvectors of $\beta$ and identifications												
PPP	$\beta_{p^{TR}}$	1	$\beta_{p^{US}}$	$\beta_{e^s}$	$\beta_{i^{TR}}$	$\beta_{i^{US}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^s}$	$\chi_{(6)}^2$
		1	-1	-1	0	0	0.022	0.015	0.007	-0.44	-	28.438
UIP	$\beta_{p^{TR}}$	0	0	0	1	-1	-0.016	-0.026	-0.005	-	-0.58	(0.000)
Standardized eigenvectors of $\beta$ and identifications												
PPP	$\beta_{p^{TR}}$	$\beta_{p^{US}}$	$\beta_{e^s}$	$\beta_{i^{TR}}$	$\beta_{i^{US}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^s}$	$\chi_{(1)}^2$	
	1	-1	-1	0.054	-0.051	-0.019	-0.007	0.007	-0.212	-	3.567	
UIP	0.282	-0.238	-0.242	1	-1	-0.016	-0.026	0.013	-	-0.147	(0.058)	

Note  $p$ -values are in parentheses

**Table 5** VECM restrictions test results and identified long-run equations with adjustment coefficients for Turkey-Euro Area

Individual exclusion of:	$H_0$	$L$ -statistics	Weak exogeneity of:	$H_0$	$L$ -statistics	Structural breaks	$H_0$	$L$ -statistics				
$p_t^{TR}$	$\beta_{p^{TR}} = 0$	15.673 (0.000)	$p_t^{TR}$	$\alpha_{p^{TR}} = 0$	14.824 (0.000)	2005:2	$\gamma_1 = \gamma_2$	20.135 (0.000)				
$p_t^{EU}$	$\beta_{p^{EU}} = 0$	10.585 (0.005)	$p_t^{EU}$	$\alpha_{p^{EU}} = 0$	4.429 (0.109)							
$e_t^e$	$\beta_{e^e} = 0$	15.821 (0.000)	$e_t^e$	$\alpha_{e^e} = 0$	12.055 (0.004)							
$i_t^{TR}$	$\beta_{i^{TR}} = 0$	8.995 (0.011)	$i_t^{TR}$	$\alpha_{i^{TR}} = 0$	5.192 (0.074)	2010:1	$\gamma_2 = \gamma_3$	28.121 (0.000)				
$i_t^{EU}$	$\beta_{i^{EU}} = 0$	7.024 (0.029)	$i_t^{EU}$	$\alpha_{i^{EU}} = 0$	4.028 (0.135)							
Standardized eigenvectors of $\beta$ and identifications												
PPP		$\beta_{p^{TR}}$	$\beta_{p^{EU}}$	$\alpha_{e^e}$	$\beta_{i^{TR}}$	$\beta_{i^{EU}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^e}$	$\chi^2_{(6)}$
	1	-1	-1	0	0	0	-0.247	0.015	-0.006	-0.197	-	27.094
UJP		$\beta_{p^{TR}}$	$\beta_{p^{EU}}$	$\alpha_{e^e}$	$\beta_{i^{TR}}$	$\beta_{i^{EU}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^e}$	$\chi^2_{(1)}$
	0	0	0	0	1	-1	-0.016	-0.026	-0.005	-	-0.014	(0.000)
Standardized eigenvectors of $\beta$ and identifications												
PPP		$\beta_{p^{TR}}$	$\beta_{p^{EU}}$	$\alpha_{e^e}$	$\beta_{i^{TR}}$	$\beta_{i^{EU}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^e}$	$\chi^2_{(1)}$
	1	-1	-1	-1	0.134	-0.135	0.006	-0.003	0.008	-0.115	-	1.219
UJP		$\beta_{p^{TR}}$	$\beta_{p^{EU}}$	$\alpha_{e^e}$	$\beta_{i^{TR}}$	$\beta_{i^{EU}}$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\alpha_{p^{TR}}$	$\alpha_{e^e}$	$\chi^2_{(1)}$
	0.551	-0.554	-0.563	1	-1	-1	0.011	0.011	0.008	-	-0.025	(0.269)

Note  $p$ -values are in parentheses

The results showed that, for both models, each parity condition is rejected when it is formulated jointly. This implies that it is possible for the adjustment in asset and commodity markets to be codependent in the international financial market. On the other hand, when each parity condition is formulated as less restrictive, not all models are rejected. This indicates that the conditions of proportionality and symmetry are met by PPP and UIP.

In a financially open economy, international parity interactions have an important implication especially for stabilization programs and exchange rate targeting. Such policies cannot be built upon the existing view that equilibrium exchange rates are based on the condition of market-clearing in Purchasing Power Parity. The exchange rates' adjustment to capital flow is because of interest rate parities, which leads targeted exchange rates to significantly deviate.

The results of this paper suggest that the question of consolidating data from tests of international parity conditions and asset mobility for emerging market economies should be given more attention. Identifying the two parity conditions might be used to assess equilibrium real exchange rate models for emerging market countries.

## References

- Dawson PJ, Sanjuan AI (2005) Structural breaks, the export enhancement program and the relationship between Canadian and the U.S. hard wheat prices. *J Agric Econ* 57:101–116
- Erlat H (2003) The nature of persistence in Turkish real exchange rates. *Emerg Mark Financ Trade* 39(2):70–97
- Harris R, Sollis R (2003) In: Chichester W (ed) *Applied time series modeling and forecasting*. Wiley, Sussex
- Hendry DF, Mizon GE (1993) Evaluating dynamic econometric models by encompassing the VAR. In: Phillips PC (ed) *Models, methods and applications of econometrics*. Basil Blackwell, Oxford, pp 272–300
- Johansen S, Juselius K (1992) Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *J Econ* 53:211–244
- Johansen S, Mosconi R, Nielsen B (2000) Cointegration analysis in the presence of structural breaks in the deterministic trend. *Econ J* 3:216–249
- Juselius K (1995) Do purchasing power parity and uncovered interest rate parity hold in the long run? An example of likelihood inference in a multivariate time-series model. *J Econ* 69:211–240
- Juselius K, MacDonald R (2004) International parity relationships between the USA and Japan. *Jpn World Econ* 16(19):17–34
- Juselius K, MacDonald R (2000) International parity relationships between Germany and the United States: a joint modelling approach. Unpublished report, European University Institute
- Lee J, Strazicich MC (2003) Minimum lagrange multiplier unit root test with two structural breaks. *Rev Econ Stat* 85:1082–1089
- Metin K (1994) A test of long-run purchasing power parity and uncovered interest parity: Turkish case. *Bilkent University Discussion Papers*, No. 94–2
- Özmen E, Gökcan A (2004) Deviations from PPP and UIP in a financially open economy: the Turkish evidence. *Appl Financ Econ* 14:779–784
- Sarno L (2000) Real exchange rate behaviour in high inflation countries: empirical evidence from Turkey, 1980–1997. *Appl Econ Lett* 7:285–291
- Telatar E, Kazdağlı H (1998) Re-examining the long-run purchasing power parity hypothesis for a high inflation country: the case of Turkey 1980–93. *Appl Econ Lett* 5:51–53

# Stochastic Volatility Models with Endogenous Breaks in Volatility Forecasting



Akram S. Hasanov  and Salokhiddin S. Avazkhodjaev 

**Abstract** The need for research on modelling and forecasting financial volatility has increased noticeably due to its essential role in portfolio and risk management, option pricing, and dynamic hedging. This paper contributes to the ongoing discussion of how researchers use regime shifts or structural breaks information to improve forecast accuracy. To accomplish this, we use the data on renewable energy markets. Thus, this study examines several models that accommodate regime shifts and investigates their forecasting performance. First, a subset of competing models (GARCH-class and stochastic volatility) employ the modified iterative cumulative sum of squares method to determine the estimation windows. This paper's novel aspect is that it studies the forecasting performance of various specifications of stochastic volatility models under this procedure. Second, we employ Markov switching GARCH models under alternative distribution assumptions. The rolling window-based forecast analysis reveals that Markov switching models offer more accurate volatility forecast results for most cases. Regarding distribution functions' relevance, the normal distribution followed by Student's  $t$ , skew Student  $t$ , and generalized hyperbolic distribution commonly dominates the series under investigation in the superior sets under all considered loss metrics.

**Keywords** Volatility modelling and forecasting · Regime shifts · Renewable energy · The rolling window

---

A. S. Hasanov (✉)

Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, Subang Jaya, 47500 Kuala Lumpur, Selangor, Malaysia

e-mail: [akram.hasanov@monash.edu](mailto:akram.hasanov@monash.edu)

S. S. Avazkhodjaev

Tashkent Institute of Finance, 60A, A. Temur Street, 100000 Tashkent, Republic of Uzbekistan

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_6](https://doi.org/10.1007/978-3-030-85254-2_6)



# 1 Introduction

The studies on modelling and forecasting financial volatility occupy a considerable portion of empirical finance literature due to its essential applications in portfolio and risk management, option pricing, and dynamic hedging. This study mainly explores the role of regime changes in forecasting the volatility in renewable energy markets through the moving window approach. Also, there have been somewhat little research works on modelling and predicting the volatility of renewable energy markets despite the substantial necessity resulting from renewable energy supporting policies worldwide. Thus, this investigation also aims to contribute to the limited literature on renewable energy markets' volatility prediction analysis. We employ the three commonly used indices in the renewable energy sector, namely, the European Renewable Energy Index (ERIX), the S&P Global Clean Energy Index (S&P GCE), and the Wilder Hill Clean Energy Index (ECO).

Past literature documents that ignoring regime changes in volatility models may overestimate the model coefficients (i.e., persistence), resulting in inaccurate volatility forecasts (see, among others, Lamoureux and Lastrapes 1990; Nomikos and Pouliasis 2011). Thus, the models that take structural breaks into account can be more suitable for predicting renewable energy market volatility.

We employ several models to accommodate structural breaks. First, some models under consideration employ the modified iterative cumulative sum of squares (ICSS) algorithm developed by Sansó et al. (2004) to determine the volatility models' estimation windows. We employ the ICSS method to determine the samples for estimations. This strategy of computing the volatility forecasts is, to some extent, similar to the approach employed in Rapach and Strauss (2008). However, while Rapach and Strauss (2008) assume Gaussian distribution for the maximum likelihood function, we consider, together with Gaussian, some other conditional densities, such as the asymmetric and symmetric fat-tailed density functions. Another novel aspect is that we examine the forecasting performance of stochastic volatility models under this procedure. Second, we employ the Markov switching GARCH models (i.e., MS-GARCH, MS-EGARCH, and MS-GJRGARCH) proposed by Haas et al. (2004) with different distributional assumptions. Besides regime shifts, we have attempted to accommodate various stylized facts in renewable energy markets such as skewness, non-normality, asymmetry, and excess kurtosis.

As usually documented, the proper conditional distribution for asset returns is essential for options valuation and asset pricing (e.g., Hasanov et al. 2018). However, the role of various conditional densities in out-of-sample forecasting analysis is still an open question in the literature. One may compute the volatility forecasts by employing various methods, depending on the variance models and conditional density functions.

Chuang et al. (2007) mentioned that a distribution function needs to comprise some essential features of asset returns, such as shapes, skewness, and kurtosis. In the literature, the statisticians have developed several complex distribution functions (e.g., Fernandez and Steel 1998; Theodossiou 1998). However, a limited number of works

focus on stochastic volatility (SV) and GARCH-type and models' volatility forecasting performance using various distribution functions (with the notable exception of Harvey et al. (1994); Giot and Laurent (2003); Chuang et al. (2007); Hasanov et al. (2018)). This paper employed the four types of conditional density functions in the empirical estimations of GARCH-type models. These are the Student  $t$  (STD), Gaussian (N), skew Student  $t$  (SSTD), and generalized hyperbolic distribution (GHYP). In comparison, we rely on two distributions for SV models, such as Gaussian and Student  $t$ .

In this study, we employ the model confidence set (MCS) test technique proposed in a study by Hansen et al. (2011) to assess the predicting performance of models under consideration following the comparatively recent research works in this context (see, among others, Charles and Darné (2017); Laporta et al. (2018); Zhang et al. (2019); and Hasanov et al. (2020)).

We contribute to a few research questions. First, we examine how one needs to consider the structural breaks to improve the forecast accuracy for the markets under consideration. Second, we investigate whether stochastic volatility models' forecasting performance is improved compared with GARCH-type models when the breaks in log-returns are considered in both models. Third, we analyze the role of distributions in volatility prediction performance. Finally, we also study whether asymmetric models perform better than symmetric models.

We organize the remainder of this study as follows. In Sect. 2, we provide some information about the data employed in this study. Section 3 comprises the methodology, including model specifications, break test, and out-of-sample forecasting analysis procedure. Section 4 includes forecasting results and provides some discussion. Finally, Sect. 5 concludes the paper and highlights some practical implications.

## 2 Data

We rely on the three most widely used stock indices in the renewable energy sector to represent the renewable energy market. First, in the analysis, we use the Wilder Hill Clean Energy Index (ECO). These series are constructed as the weighted mean of the corporate stocks of publicly traded entities whose business operations may have benefited noticeably from a conventional societal position towards cleaner energy use and conservation. Second, we selected the European Renewable Energy Index (ERIX). This index series comprises 10 renewable energy companies' corporate stocks in biomass, water, wind, and solar energy in Europe. Finally, we took the S&P Global Clean Energy Index (S&P GCE). These index series are computed as the weighted mean of the corporate shares of 30 companies devoted to developing renewable energy technologies worldwide. The sample periods for all three indices end on November 30, 2020, and start on January 3, 2005. We retrieve the data on these renewable energy indices from the Bloomberg database.

### 3 Empirical Models

We use some univariate models to predict renewable energy market volatility. We employ the historical mean (HM), exponentially smoothing (ES), GARCH-class, Markov-switching GARCH (MS-GARCH), and stochastic volatility models.<sup>1</sup> We take the structural breaks or regime shifts into account by using two ways. First, we rely on the adjusted ICSS method to select the estimation windows for different specification of SV and GARCH-type conditional variance models. Second, we employ Markov-switching GARCH models under various distribution assumptions to model and forecast conditional volatility.

#### 3.1 The Stochastic Volatility Models

The main characteristics of the SV model are its stochastic and time-varying features of the variance evolution. Specifically, one assumes that the log-variance of the model follows an AR(1) process. Hence, we may specify the following models. Let vector  $y = (y_1, \dots, y_n)^T$  comprises the demeaned log-return observations of an asset. The usual SV model assuming normal distribution can be specified as:

$$\begin{aligned} y_t &= \mathbf{x}_t \boldsymbol{\beta} + \exp\left(\frac{h_t}{2}\right) \varepsilon_t \\ h_{t+1} &= \mu + \varphi(h_t - \mu) + \sigma \eta_t \\ \varepsilon_t &\sim \text{N}(0, 1) \\ \eta_t &\sim \text{N}(0, 1) \end{aligned} \tag{1}$$

We have also considered the SV model with the conditional Student's  $t$  distribution suggested by Harvey et al. (1994).

$$\begin{aligned} y_t &= \mathbf{x}_t \boldsymbol{\beta} + \exp\left(\frac{h_t}{2}\right) \varepsilon_t \\ h_{t+1} &= \mu + \varphi(h_t - \mu) + \sigma \eta_t \\ \varepsilon_t &\sim t_v(0, 1) \\ \eta_t &\sim \text{N}(0, 1) \end{aligned} \tag{2}$$

where  $t_v$  is the Student's  $t$  distribution with  $v$  degrees of freedom, unit variance and zero mean.

We also employ the SV model that accounts for the leverage effect.

---

<sup>1</sup> For a detailed description of the HM and ES models, we refer interested readers to Sadorsky (2006) and Hasanov et al. (2020).

$$\begin{aligned}
 y_t &= \mathbf{x}_t \boldsymbol{\beta} + \exp\left(\frac{h_t}{2}\right) \varepsilon_t \\
 h_{t+1} &= \mu + \varphi(h_t - \mu) + \sigma \eta_t \\
 \varepsilon_t &\sim \mathbf{N}(0, 1) \\
 \eta_t &\sim \mathbf{N}(0, 1)
 \end{aligned}
 \tag{3}$$

where the correlation matrix of  $(\varepsilon_t, \eta_t)$  is defined as

$$\Sigma^\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

In the above SV models,  $\varepsilon_t$  and  $\eta_t$  are independent. The  $h$  denotes the log-variance process,  $\mathbf{x}_t$  is a vector of regressors, and  $\boldsymbol{\beta}$  is a vector of coefficients. We use the Greek letters  $\mu$ ,  $\varphi$ , and  $\sigma$  to denote the parameters of models.

### 3.2 The GARCH-Type Models

This study’s first deterministic conditional volatility model is the simple GARCH(1, 1) model proposed by Bollerslev (1986). The GARCH(1, 1) model for a given log-return series (i.e.,  $r_t$ ) with a mean model, which follows ARMA( $p, q$ ) process can be written as:

$$r_t = \mu + \sum_{i=1}^p \phi_i r_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}
 \tag{4}$$

$$\varepsilon_t = \sigma_t \varepsilon_t, \sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2
 \tag{5}$$

where  $\varepsilon_t$  is a series of identically and independently distributed (i.i.d.) random variables with unit variance and zero mean;  $\sigma_t^2$  denotes the conditional variance, and the parameters of mean and variance models are as follows:  $|\delta| < 1$ ,  $\omega > 0$ ,  $\alpha_1 \geq 0$ ,  $\beta_1 \geq 0$ , and  $\alpha_1 + \beta_1 < 1$ .

As commonly noted, the financial markets frequently demonstrate evidence of the asymmetric volatility phenomenon. This implies that research works should utilize the models that encompass the data’s asymmetry features to examine financial markets’ volatility. The GJR-GARCH specification suggested in Glosten et al. (1993) is another GARCH-class of specification that accounts for the phenomenon of asymmetry. The mean and variance models can be specified as:

$$r_t = \mu + \sum_{i=1}^p \phi_i r_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}
 \tag{6}$$

$$\sigma_t^2 = \omega + \{\alpha_1 + \gamma_1 I(\varepsilon_{t-1} > 0)\} \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (7)$$

where  $\omega > 0$ ,  $\alpha_1 \geq 0$ ,  $\alpha_1 + \gamma_1 \geq 0$ ,  $\beta_1 \geq 0$ , and  $I(\varepsilon_{t-1} > 0)$  is an indicator variable that takes one when  $\varepsilon_{t-1} < 0$  and obtains zero if the argument is false. One may capture an asymmetric impact in the log-return series by analyzing the coefficient estimate of  $\gamma_1$ .

To incorporate conditional variances' asymmetric responses to negative and positive shocks with a similar absolute value, Nelson (1991) developed the exponential GARCH (EGARCH) model specified as:

$$\ln(\sigma_t^2) = \omega + \alpha |z_{t-1}| + \gamma z_{t-1} + \beta \ln(\sigma_{t-1}^2) \quad (8)$$

where  $z_{t-1} = \varepsilon_{t-1} \sigma_{t-1}^{-1}$ .

The GARCH and GJR-GARCH models impose non-negative constraints on the variance equation parameters, while there are no restrictions imposed on the variance coefficients in the EGARCH model.

This paper considers four conditional distributions in the empirical estimations of GARCH-type models. These distributions are as follows: the standardized Student  $t$  (STD), standard Gaussian (N), skew standardized Student  $t$  (SSTD), and generalized hyperbolic distribution (GHYP). For a detailed description of the functional forms of the aforementioned conditional distribution functions, we refer interested readers to Hasanov et al. (2018) and their paper's references.

### 3.3 *The SV and GARCH-Type Models with Endogenously Determined Breaks*

To accommodate the possible structural changes in the log-return series, we rely on the adjusted ICSS procedure developed in the paper by Sansó et al. (2004) to select the estimation windows for the SV and GARCH-type conditional variance model specifications. In the initial step, we apply the  $\kappa_2$  test to all existing observations one through  $T$ . Assume we detected one or more breaks through the modified ICSS procedure. The last break has been found to happen at the time  $T_B$ . Then all models considered in this study are estimated using log-return series starting from  $T_B + 1$  to  $T$  to get an estimate of  $\sigma_{T+1}^2$ . Then we again apply the  $\kappa_2$  test to observations from two through  $T + 1$ . The last breakpoint is now found to happen at the time  $T_{1B}$ . In the next step, we estimate all models for log-returns from  $T_{1B} + 1$  to  $T + 1$  to compute the following estimate of  $\sigma_{T+1}^2$ . We continue this process until all out-of-sample forecast period is exhausted. In sum, we rely on the  $\kappa_2$  break detection test to find the estimation sample so-called the "structural break" model.<sup>2</sup>

<sup>2</sup> In Tables 1, 2 and 3, we use the prefix "SB" to indicate the models with structural breaks.

**Table 1** Model confidence set for ECO ( $s = 1, R = 500$ )

AE	HMAE		HMSE		QLIKE	
	1.000	1.000	1.000	1.000	1.000	1.000
SB-HM		MS-GARCH-STD	MS-GARCH-STD	MS-GARCH-STD	MS-GJR-GARCH-SSTD	1.000
	0.805	MS-GARCH-N	MS-GARCH-N	MS-GARCH-N	MS-GJR-GARCH-N	0.448
	0.805	MS-GARCH-SSTD	MS-GARCH-SSTD	MS-GARCH-SSTD	MS-GARCH-STD	0.963
	0.582	MS-GJR-GARCH-N	MS-GJR-GARCH-N	MS-GJR-GARCH-N	MS-EGARCH-SSTD	0.937
	0.582	MS-GJR-GARCH-SSTD	MS-GJR-GARCH-SSTD	MS-GJR-GARCH-SSTD	MS-GARCH-N	0.937
	0.488	MS-EGARCH-SSTD	MS-EGARCH-SSTD	MS-EGARCH-SSTD	MS-GARCH-SSTD	0.931
		MS-GJR-GARCH-STD	MS-GJR-GARCH-STD	MS-GJR-GARCH-STD	MS-GJR-GARCH-STD	0.519
		MS-EGARCH-N	MS-EGARCH-STD	MS-EGARCH-STD	MS-EGARCH-N	0.324
			MS-EGARCH-N	MS-EGARCH-N	MS-EGARCH-STD	0.324
			SB-ES	SB-ES	SB-ES	0.324
			SB-GJR-GARCH-SSTD	SB-GJR-GARCH-SSTD	SB-GJR-GARCH-STD	0.324
			SB-GJR-GARCH-GHYP	SB-GJR-GARCH-GHYP	SB-GJR-GARCH-SSTD	0.324
			SB-GJR-GARCH-STD	SB-GJR-GARCH-STD	SB-HM	0.324
			SB-GJR-GARCH-N	SB-GJR-GARCH-N	SB-GJR-GARCH-GHYP	0.324
			SB-GARCH-SSTD	SB-GARCH-SSTD	SB-GJR-GARCH-N	0.324
			SB-GARCH-GHYP	SB-GARCH-GHYP	SB-GARCH-STD	0.324
			SB-GARCH-STD	SB-GARCH-STD	SB-GARCH-SSTD	0.324
			SB-HM	SB-HM	SB-GARCH-GHYP	0.324
			SB-GARCH-N	SB-GARCH-N	SB-GARCH-N	0.324

(continued)

Table 1 (continued)

AE	HMAE		HMSE		QLIKE	
SB-HM	1.000	MS-GARCH-STD	1.000	MS-GARCH-STD	MS-GJR-GARCH-SSTD	1.000
				SB-SV-leverage	SB-SV-leverage	0.302
				SB-SV	SB-SV	0.154
				SB-SV-t	SB-EGARCH-N	0.154
				SB-EGARCH-N	SB-SV-t	0.137

Notes: The table includes the Superior set of models and the corresponding  $p$ -values calculated through  $T_R$  statistics under the considered loss metrics with  $R = 500$  and  $s = 1$ . The prefix "SB" shows the models with structural changes (endogenously detected). The distributions employed in the maximum likelihood estimation process are (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) a skew Student  $t$  (SSTD), and (iv) a generalized hyperbolic (GHYP). The adjusted ICSS regime change test is employed to detect an estimation sample in the models with structural breaks. The prefix "MS" indicates the Markov switching GARCH models, which are estimated by using three conditional distributions: (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) and a skew Student  $t$  distribution (SSTD)

**Table 2** Model confidence set for ERIX ( $s = 1, R = 500$ )

AE	HMAE		HMSE		QLIKE	
	1.000	MS-GARCH-N	1.000	MS-GARCH-N	1.000	MS-EGARCH-SSTD
MS-EGARCH-SSTD	1.000	MS-GARCH-N	1.000	MS-GARCH-N	1.000	MS-EGARCH-SSTD
MS-EGARCH-N	0.629	MS-GARCH-SSTD	0.832	MS-GARCH-STD	0.646	MS-GARCH-N
MS-GJR-GARCH-SSTD	0.629	MS-GARCH-STD	0.829	MS-EGARCH-N	0.646	MS-EGARCH-STD
MS-GARCH-N	0.629	MS-EGARCH-SSTD	0.829	MS-GARCH-SSTD	0.646	MS-GARCH-SSTD
MS-EGARCH-STD	0.629	MS-GJR-GARCH-SSTD	0.820	MS-EGARCH-SSTD	0.646	MS-EGARCH-N
MS-GARCH-STD	0.629	MS-GJR-GARCH-N	0.439	MS-GJR-GARCH-SSTD	0.608	MS-GJR-GARCH-SSTD
SB-SV-leverage	0.629	MS-EGARCH-N	0.439	MS-EGARCH-STD	0.435	MS-GARCH-STD
MS-GARCH-SSTD	0.507	MS-GJR-GARCH-STD	0.439	MS-GJR-GARCH-N	0.435	MS-GJR-GARCH-STD
MS-GJR-GARCH-N	0.446	MS-EGARCH-STD	0.439	MS-GJR-GARCH-STD	0.435	MS-GJR-GARCH-N
SB-GJR-GARCH-SSTD	0.446	SB-ES	0.439	SB-ES	0.435	SB-SV-leverage
SB-GJR-GARCH-STD	0.332	SB-SV	0.439	SB-GARCH-N	0.435	SB-HM
		SB-GARCH-N	0.439	SB-GARCH-GHYP	0.435	SB-GJR-GARCH-STD
		SB-SV-t	0.439	SB-GARCH-STD	0.435	SB-GJR-GARCH-SSTD
		SB-GARCH-STD	0.439	SB-GARCH-SSTD	0.435	SB-EGARCH-SSTD
		SB-GARCH-SSTD	0.439	SB-SV	0.435	SB-GJR-GARCH-GHYP
		SB-GARCH-GHYP	0.439	SB-SV-t	0.435	SB-EGARCH-GHYP
		SB-HM	0.439	SB-EGARCH-N	0.435	SB-SV-t
		MS-EGARCH-N	0.439	SB-HM	0.435	SB-SV
		SB-SV-leverage	0.439	SB-EGARCH-SSTD	0.435	SB-GJR-GARCH-N
				SB-EGARCH-GHYP	0.364	SB-EGARCH-N
				SB-GJR-GARCH-N	0.364	SB-GARCH-SSTD

(continued)



**Table 2** (continued)

AE	HMAE		HMSE		QLIKE		
MS-EGARCH-SSTD	<b>1.000</b>	MS-GARCH-N	<b>1.000</b>	MS-GARCH-N	<b>1.000</b>	MS-EGARCH-SSTD	<b>1.000</b>
				SB-GJR-GARCH-STD	0.253	SB-GARCH-STD	0.217
				SB-GJR-GARCH-GHYP	0.253	SB-GARCH-GHYP	0.217
				SB-GJR-GARCH-SSTD	0.149	SB-GARCH-N	0.217
				SB-SV-leverage	0.149	SB-ES	0.122

*Notes:* The table includes the Superior set of models and the corresponding  $p$ -values calculated through  $T_R$  statistics under the considered loss metrics with  $R = 500$  and  $s = 1$ . The prefix “SB” shows the models with structural changes (endogenously detected). The distributions employed in the maximum likelihood estimation process are (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) a skew Student  $t$  (SSTD), and (iv) a generalized hyperbolic (GHYP). The adjusted ICSS regime change test is employed to detect an estimation sample in the models with structural breaks. The prefix “MS” indicates the Markov switching GARCH models, which are estimated by using three conditional distributions: (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) and a skew Student  $t$  distribution (SSTD)

**Table 3** Model confidence set for SPGCE ( $s = 1, R = 500$ )

AE	HMAE		HMSE		QLIKE	
	1.000	MS-GARCH-N	1.000	MS-GARCH-N	1.000	MS-GARCH-N
SB-SV-t	0.629	MS-GARCH-STD	0.832	MS-GARCH-SSTD	0.786	MS-GJR-GARCH-N
	0.629	MS-GARCH-SSTD	0.829	MS-GARCH-STD	0.704	MS-EGARCH-SSTD
	0.629	MS-GJR-GARCH-N	0.829	MS-GJR-GARCH-N	0.704	MS-EGARCH-STD
	0.629		0.820	MS-EGARCH-SSTD	0.226	MS-GARCH-STD
	0.629		0.439	MS-GJR-GARCH-STD	0.226	MS-GARCH-SSTD
	0.629		0.439	MS-EGARCH-N	0.226	MS-GJR-GARCH-STD
	0.507		0.439	MS-GJR-GARCH-SSTD	0.226	MS-EGARCH-N
	0.446		0.439	MS-EGARCH-STD	0.226	MS-GJR-GARCH-SSTD
	0.446		0.439	SB-HM	0.226	SB-HM
	0.332		0.439	SB-EGARCH-N	0.226	SB-EGARCH-N
			0.439	SB-EGARCH-STD	0.226	SB-EGARCH-STD
			0.439	SB-EGARCH-SSTD	0.226	SB-EGARCH-GHYP
			0.439	SB-EGARCH-GHYP	0.226	SB-EGARCH-SSTD
			0.439	SB-GJR-GARCH-STD	0.226	SB-GJR-GARCH-N
			0.439	SB-GJR-GARCH-GHYP	0.226	SB-SV-leverage
			0.439	SB-GJR-GARCH-SSTD	0.226	SB-GJR-GARCH-GHYP
			0.439	SB-GJR-GARCH-N	0.226	SB-GARCH-GHYP
			0.439	SB-SV-leverage	0.226	SB-GJR-GARCH-STD
				SB-SV	0.226	SB-GARCH-SSTD
				SB-SV-t	0.226	SB-GARCH-N

(continued)

**Table 3** (continued)

AE	HMAE		HMSE		QLIKE		
SB-SV-t	<b>1.000</b>	MS-GARCH-N	<b>1.000</b>	MS-GARCH-N	<b>1.000</b>	MS-GARCH-N	<b>1.000</b>
				SB-GARCH-N	0.226	SB-GJR-GARCH-SSTD	0.180
				SB-GARCH-GHYP	0.226	SB-GARCH-STD	0.180
				SB-GARCH-STD	0.226	SB-SV	0.154
				SB-GARCH-SSTD	0.226	SB-SV-t	0.154

*Notes:* The table includes the Superior set of models and the corresponding  $p$ -values calculated through  $T_R$  statistics under the considered loss metrics with  $R = 500$  and  $s = 1$ . The prefix “SB” shows the models with structural changes (endogenously detected). The distributions employed in the maximum likelihood estimation process are (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) a skew Student  $t$  (SSTD), and (iv) a generalized hyperbolic (GHYP). The adjusted ICSS regime change test is employed to detect an estimation sample in the models with structural breaks. The prefix “MS” indicates the Markov switching GARCH models, which are estimated by using three conditional distributions: (i) a standardized Student  $t$  (STD), (ii) a Gaussian (N), (iii) and a skew Student  $t$  distribution (SSTD)

If regime shifts or structural changes occur in the log-return series, employing the whole available log-return series to estimate a model may generate inaccurate forecasts, despite having a lesser variance (Pesaran and Timmermann 2007). The existing theory in this context advocates that neglecting the structural breaks or regime shifts may induce upward biases in volatility persistence estimates (for example, Lamoureux and Lastrapes 1990; Mikosch and Stărică 2004). Hence, the observations only over after the break period have been used (i.e., from  $T_B + 1$  to  $T$ ) to estimate the models with breaks, given after-break period is sufficient to run the estimations. Rapach and Strauss (2008) mentioned that a possible shortcoming of this method is that the observations might be insufficient to estimate the model coefficients.

### 3.4 Markov-Switching GARCH Models

In addition to the models described in the previous sections, we also estimate the MS-GARCH model proposed by Haas et al. (2004), assuming different distribution assumptions (Student  $t$ , Gaussian, and skew Student  $t$ ). Ardia et al. (2018) specify the MS-GARCH model as:

$$r_t | (s_t = k, \Psi_{t-1}) \sim D(0, h_{k,t}, \vartheta_k) \quad (9)$$

where  $D(0, h_{k,t}, \vartheta_k)$  is a distribution (i.e., continuous) function with a mean equal to zero, and a time-varying conditional volatility  $h_{k,t}$  in regime  $k$ . Here, a vector,  $\vartheta_k$ , includes additional parameters like skew or tail parameter. The symbol,  $\Psi_{t-1}$ , denotes the available information set. Here, the state variable,  $s_t$ , changes in line with a first-order homogeneous Markov chain with  $k$  states. We rely on three conditional variance models: the GARCH, the EGARCH, and the GJR-GARCH.

## 4 Results and Discussion

The coefficient estimates of all considered models<sup>3</sup> outlined in Sect. 3 have been used to produce daily single-step ahead predictions for conditional variances. We move forward the beginning as well as end dates of the estimation period one day. We re-estimate the model coefficients, and finally, obtained new parameter estimates are employed to predict daily single-step variance (i.e., conditional) over the pre-set out-of-sample period. This process continues until the out-of-sample period has been completed. The numerous studies in the literature applied a rolling window

---

<sup>3</sup> Since HM estimation relies on non-parametric way of estimating the volatility, we do not have any parameters to estimate in HM.

approach of computing out-of-sample predictions. (e.g., Sadorsky 2006; Wen et al. 2016; Charles and Darné 2017; Hasanov et al. 2018; Hasanov et al. 2020).

Following many previous studies, we separate the entire sample for the log-return series under consideration into two parts: in-sample and out-of-sample. While the out-of-sample period comprises the last  $R$  observations, the in-sample period covers the initial  $T$  observations. In this study, the out-of-sample period  $R$  is set to include 500 observations.

As mentioned earlier, we rely on the MCS testing procedure in forecast comparison analysis. We predict the volatility employing the models defined in the previous section estimated on each of the three renewable energy market returns to apply this technique. The MCS testing procedure has been shown to offer a robust comparison for the predictions generated from many models simultaneously and generate a superior set that includes the superior models in forecasting performance with the given pre-specified confidence level. In this study, we have set the confidence level for the MCS test to  $\alpha = 0.90$ . The software code we have written for the analysis relies on the R (version 4.0.2) software and the MCS (Catania and Bernardi 2015), the rugarch (Ghalanos 2016), the stochvol (Hosszejni and Kastner 2016), and the MSGARCH (Ardia et al. 2019) packages.

In this paper, the MCS test is based on the following loss metrics:

$$\begin{aligned}
 AE_{t+1} &= \left| \tilde{V}_{t+1} - \hat{V}_{t+1} \right| \\
 HMAE_{t+1} &= \left| 1 - \frac{\tilde{V}_{t+1}}{\hat{V}_{t+1}} \right| \\
 HMSE_{t+1} &= \left( 1 - \frac{\tilde{V}_{t+1}}{\hat{V}_{t+1}} \right)^2 \\
 QLIKE_{t+1} &= \log\left(\hat{V}_{t+1}^2\right) + \tilde{V}_{t+1}^2 \hat{V}_{t+1}^{-2}
 \end{aligned}$$

where  $\hat{V}_{t+1}$  is the volatility forecasts computed by the estimated models at time  $t$ . And,  $\tilde{V}_{t+1}$  is an actual volatility's proxy at time  $t$ . The squared returns serve as a proxy (see Sadorsky 2006; Hasanov et al. 2020 among many others).

Table 1 shows that the majority of Markov switching models are in the superior set  $\hat{M}_{0.9}$  in terms of three loss metrics (i.e., HMAE, HMSE, and QLIKE). Thus, these three Markov switching models happen to be the most accurate single-period forecasting models for the ECO series. Moreover, the SB-HM model is also in the superior set of models  $\hat{M}_{0.9}$  under most of the loss metrics, and, therefore, this is also considered as a promising model. Under  $AE_1$  loss function, the MCS procedure excludes all other competing models except SB-HM from a superior set of models (SSM). The model confidence set comprises the GARCH-type models and SV models

with endogenously determined structural breaks under two out of four loss criteria (i.e., HMSE and QLIKE).

As one can see from Table 2, under the AE and QLIKE, the MS-EGARCH model with SSTD distribution for the ERIX series shows the highest forecasting performance according to the  $p$ -values computed by the MCS algorithm. Meanwhile, the MS-GARCH with Gaussian distribution provides the most accurate forecasts under HMAE and HMSE loss criteria. The SB-SV model with leverage effects survives in all SSMs, indicating that this is also a favourable model. The MCS procedure selects SB-GJR-GARCH with Student  $t$ , and skew Student  $t$  distributions in the SSM under AE, HMSE, and QLIKE loss functions.

The three loss criteria (HMAE, HMSE, and QLIKE) choose the MS-GARCH-N model as the most promising forecasting model for the SPGCE series. Meanwhile, the MS-GARCH with SSTD, MS-GARCH with STD, and MS-GJR-GARCH with N survive in three out of four SSMs, showing that they are also favourable models. Moreover, the SB-SV-t model for the SPGCE series turned out to be the model with the highest forecasting performance under the AE criterion, based on the  $p$ -values computed by the MCS test (see Table 3). This model also appears in the superior set of models  $\hat{M}_{0,9}$  under HMSE and QLIKE loss criteria. The rest of the models under consideration for the SPGCE index survive when the MCS test relies on HMSE and QLIKE forecast summary statistics.

We also looked into the importance of distributions in forecasting analysis for the models under consideration. The findings for the relevance of distributions in single-step-ahead forecasting are rather mixed. The results show that distribution specifications employed in rolling estimations of models with the most accurate forecasting performance are not consistent with the true underlying distribution of returns. Besides, no dominant distribution is found for the markets under study, which increases forecasting performance. In general, the normal distribution followed by Student  $t$ , skew Student  $t$ , and generalized hyperbolic distribution commonly dominates for the series under investigation in the model confidence sets under all considered accuracy criteria. This result is consistent with those found by Chuang et al. (2007) and Hasanov et al. (2018), who find that the distribution function's complexity does not consistently outperform the less complex one because of the over-fitting problem.

It is worth mentioning that most models in SSMs are asymmetric models across all series and loss metrics. Thus, the renewable energy market participants must not neglect the stylized facts like regime shifts or structural changes and asymmetry when they perform risk management. Moreover, we analyze whether GARCH-based and SV models' forecasting performance, which considers the log-return series' regime changes, is improved. The results suggest that SB-GARCH and SB-SV models show a similar forecasting performance across all the markets under consideration.

## 5 Conclusion and Implications

In this study, we have addressed several research questions. We investigate how one needs to use the structural breaks information to improve the forecast accuracy for the markets under consideration. We find that the Markov switching GARCH models are the superior one-period forecasting models for all markets under investigation. Moreover, we analyze whether stochastic volatility models' forecasting performance is improved compared with GARCH-type models when the endogenously determined breaks in log-returns are considered in both models. The results suggest that both GARCH-type and SV models show a similar forecasting performance across all the markets under consideration.

Also, we analyze the relevance of several distribution specifications in volatility forecasting accuracy analysis. The results indicate no dominant asymmetric and fat-tailed distribution for the markets under study, which increases forecasting performance. In general, the normal distribution followed by Student  $t$ , skew Student  $t$ , and generalized hyperbolic distribution commonly dominates for the series under investigation in the model confidence sets under all considered accuracy criteria. Therefore, renewable energy investors and policymakers must be cautious when employing the SV and GARCH-type models to forecast market volatility. The models with complex distribution functions do not necessarily lead to better forecasting results. Finally, we also study the relevance of typical renewable energy markets like asymmetry. The results suggest that most models in superior sets are asymmetric models across all series and loss measures. Thus, renewable energy markets' participants must not neglect the essential stylized facts when they perform risk management.

This paper's modelling approach helps investors or market participants identify future renewable energy market fluctuations. The market participants, such as portfolio managers and international investors, can predict the future renewable energy market dynamics and design proper portfolio selection and risk management. They intend to generate a more accurate return and volatility predictions to evaluate the portfolio risk exposure and update the hedge ratio according to computed predictions. As Kilian and Park (2009) note, the hedge ratio's continuous adjustment confirms the proper dynamic hedging strategies.

**Acknowledgements** The first author acknowledges the financial support provided by Monash University Malaysia under the Seed grant (B-2-2020).

## References

- Ardia D, Bluteau K, Boudt K, Catania L (2018) Forecasting risk with Markov-switching GARCH models: a large-scale performance study. *Int J Forecast* 34:733–747
- Ardia D, Bluteau K, Boudt K, Catania L, Ghalanos A, Peterson B, Trottier D-A (2019) Package 'MSGARCH'. <https://cran.r-project.org>
- Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Econ* 31:307–327

- Catania L, Bernardi M (2015) Package 'MCS'. <https://cran.r-project.org>
- Charles A, Darné O (2017) Forecasting crude-oil market volatility: further evidence with jumps. *Energy Econ* 67:508–519
- Chuang I, Lu J, Lee P (2007) Forecasting volatility in the financial markets: a comparison of alternative distributional assumptions. *Appl Financ Econ* 17:1051–1060
- Fernandez C, Steel MF (1998) On Bayesian modeling of fat tails and skewness. *J Am Stat Assoc* 93:359–371
- Ghalanos A (2016) Rugarch: univariate GARCH models. R package version 1.4-4. <https://cran.r-project.org>
- Giot P, Laurent S (2003) Market risk in commodity markets: a VaR approach. *Energy Econ* 25:435–457
- Glosten LR, Jaganathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess returns on stocks. *J Financ* 48:1779–1801
- Haas M, Mittnik S, Paolella MS (2004) A new approach to Markov-switching GARCH models. *J Financ Economet* 2(4):493–530
- Hansen PR, Lunde A, Nason JM (2011) The model confidence set. *Econometrica* 79:453–497
- Harvey AC, Ruiz E, Shephard N (1994) Multivariate stochastic variance models. *Rev Econ Stud* 61(2):247–264
- Hasanov AS, Poon WC, Al-Freedi A, Heng ZY (2018) Forecasting volatility in the biofuel feedstock markets in the presence of structural breaks: a comparison of alternative distribution functions. *Energy Econ* 70:307–333
- Hasanov AS, Shaiban MS, Al-Freedi A (2020) Forecasting volatility in the petroleum futures markets: a re-examination and extension. *Energy Econ* 86:104626
- Hosszejni D, Kastner G (2016) Package 'stochvol'. <https://cran.r-project.org>
- Kilian L, Park C (2009) The impact of oil price shocks on the US stock market. *Int Econ Rev* 50:1267–1287
- Lamoureux CG, Lastrapes WD (1990) Persistence in variance, structural change, and the GARCH model. *J Bus Econ Stat* 8(2):225–234
- Laporta AG, Merlo L, Petrella L (2018) Selection of value at risk models for energy commodities. *Energy Econ* 74:628–643
- Mikosch T, Střaričá C (2004) Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev Econ Stat* 86:378–390
- Nelson D (1991) Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59:347–370
- Nomikos NK, Pouliasis PK (2011) Forecasting petroleum futures markets volatility: the role of regimes and market conditions. *Energy Econ* 33:321–337
- Pesaran MH, Timmermann A (2007) Selection of estimation window in the presence of structural breaks. *J Econ* 137:134–161
- Rapach DE, Strauss JK (2008) Structural breaks and GARCH models of exchange rate volatility. *J Appl Econ* 23:65–90
- Sadorsky P (2006) Modeling and forecasting petroleum futures volatility. *Energy Econ* 28:467–488
- Sansó A, Arragó V, Carrion JL (2004) Testing for change in the unconditional variance of financial time series. *Rev Econ Financ* 4:32–53
- Theodossiou P (1998) Financial data and the skewed generalised-t distribution. *Manag Sci* 44:1650–1661
- Wen F, Gong X, Cai S (2016) Forecasting the volatility of crude oil futures using HARtype models with structural breaks. *Energy Econ* 59:400–413
- Zhang Y-J, Yao T, Ling-Yun He L-Y, Ripple R (2019) Volatility forecasting of crude oil market: can the regime switching GARCH model beat the single-regime GARCH models? *Int Rev Econ Financ* 59:302–317



# Effect in Quality Control Based of Hotelling $T^2$ and CUSUM Control Chart



Hakan Eygü 

**Abstract** Today, quality control methods have been used quite extensively. In statistical surveys, the measurements of sampling units according to the variable under consideration are expensive in all sense and based on probability random sampling units according to same variable by means of a method which is not expensive at all. In this study, the researcher created Hotelling's  $T^2$  control charts, a multivariate statistical process control method. The performances of simple random sampling and ranked set sampling methods were compared to one another using these control charts. A statistics program was performed to see the average run length values for the comparison of the sampling performances. As a result of the study, it is determined that the process is examined by statistical quality control methods rather than bivariate methods when there is a relationship among variables in the processes covering more than one quality variable. Further research in the RSS method proved to be more efficient when units are difficult and costly to measure.

**Keywords** Statistical quality control · Control charts · Average run length · Sampling method

## 1 Introduction

Production can be defined as manufacturing a physical good or service to fulfill a requirement. Quality of the goods or services created in this process has received significant attention since earlier ages. The long lasting efforts to produce goods in the best conditions surely aimed to increase the quality level. Manufacturing goods or services occur in a process. Variability of manufacturing process eventually determines the quality of goods or services. Variability of a manufacturing process can be identified according to several measurable specifications of that process. Aim of a quality assessment application is to follow the process in the production line

---

H. Eygü (✉)

Faculty of Economics and Administrative Sciences, Department of Econometrics, Statistical Research, Ataturk University, Erzurum, Turkey  
e-mail: [hakaneygu@atauni.edu.tr](mailto:hakaneygu@atauni.edu.tr)

while it is working. It is almost impossible to measure all produced items. Instead, small quantity of samples is selected in order to measure and reach conclusions about behaviors of the process by drawing charts which are showing the changes in time. Statistics is important in quality management as it increase the understanding of variability and establishment possibilities. During the investigation of manufacturing process, one of the most important tools to understand variability and establishment is statistical quality control diagrams.

Statistics is a collection of techniques useful for making decisions about a process or population based on an analysis of the information contained in a sample from that population. Statistical methods play a vital role in quality improvement. Today, international companies aim to produce high quality products and to present their products to the market on time. The first aim of these companies should be to produce good service that can be sold and can work. Nevertheless they do offer some insight into the true nature of “quality” because they focus directly on the respondent rather than on a thing or an action being judged. In a popular sense, identifying quality is purely a judgment call. It is entirely dependent on the perceptions of the individuals or collection of individuals, making the determination. In the other sense, it refers to a body of methods by which useful conclusions can be drawn from numerical data. Thus one may say statistics is based in large part on the law of large numbers and the mathematical theory of probability.

It can be said that concept of quality is the most important characteristic of any product. Due to this feature, the development in the field of quality control continues in the first quarter of the twenty-first century. Because consumers’ awareness is increasing, it shows that consumer demands are inevitable. This situation has increased competition among firms. Manufacturers have begun to produce high quality products at minimum cost. Because manufacturing company government procurement agency or other organization in which substantial statistical quality control applications are to be made, experience indicates that appropriately there may be four levels of understanding of the subject. The first of these is the mathematics on which are based statistical techniques, and the second level is the various types of control charts and sampling tables. A third level is that of a broad understanding of the objectives and statistical quality control. This type of understanding is particularly helpful at higher management levels. Finally, the fourth level calls merely for use of one or more of the techniques on a rule of thumb basis.

It is inevitable to measure the quality of products in many industries. For this reason, researches have been carried out on the development of methods related to this subject. Statistical process control methods allow a product to be produced to meet the most economic and needs. For this purpose the collected data are used at all stages of production using statistical techniques. In this phase, control methods help Statistical Process Control (SPC) practitioners to identify the time of a change after a control chart generates a signal. Using change point estimation with the monitoring system, we have assumed that the researcher is knowledgeable about univariate statistical estimation and control procedures (such as Shewhart charts).

In this study, the performances of simple random sampling (SRS) and ranked set sampling (RSS) methods were compared to one another using these control charts.

Furthermore, these sampling methods are applied to the Hotelling  $T^2$  and CUSUM control chart. These quality control methods for obtaining ranked set samples are described, and the structural differences between ranked set samples and simple random samples are discussed. Throughout this study we are assuming that the underlying distribution is normal.

## 2 Methodology

### 2.1 Statistical Quality Control Charts

Statistical process control is a powerful collection of problem-solving tools useful in achieving process stability and improving capability through the reduction of variability. Sometimes an important product of statistical quality control may be the establishment of effective process inspection where none has previously existed. In some manufacturing concerns there is little or no process inspection. Although the introduction of process inspection is sometimes a product, it should be noted that a direct object of statistical quality control is to provide a new tool that makes process inspection more effective. Information obtained by the process investigation by checker or by machine operators often incorrectly used to make machine adjustments.

Montgomery (2009) stated that statistical quality control is one of greatest technological developments of the twentieth century because it is based on sound underlying principles, is easy to use, has significant impact, and can be applied to any process. This method techniques industry to develop and monitor processes is used various control charts have been developed to monitor the variables in the process and to detect uncontrolled conditions that degrade the quality of the products (Noorossana and Vaghefi 2006). These control cards give a graphical view. The chart contains a center line (CL) that represents the average value of the quality characteristic corresponding to the in-control state. Two other horizontal line, called the upper control limit (UCL) and the lower control limit (LCL), are also shown in the chart. The process may be outside these control limits. Shewhart (1931) has pointed out so clearly, if we fail to control the process in the statistical sense, the process is not performing in an economical fashion. Researchers are spending more money to make the product than they would if the process in control. These control limits are chosen so that if the process is in control, nearly all of the sample points will fall between them. As long as the points plot within the control limits, the process is assumed to be in control, and no action is necessary. However, a point that plots outside of the control limits is interpreted as evidence that the process is out of control, and investigation and corrective action are required to find and eliminate the assignable cause or causes responsible for this behavior. It is customary to connect the sample points on the control chart with straight-line segments so that it is easier to visualize how the sequence of points has evolved over time (Montgomery 2009). MacGregor and

Kourti (1995) suggested that control limits narrower than three-sigma be used in the chart for individuals to enhance its ability to detect small process shifts. Even if all the points plot inside the control limits, if they behave in a systematic or nonrandom manner, then this could be an indication that the process is out of control. We may give a general model for a control chart. Let  $w$  be a sample statistic that measures some quality characteristic of interest, and suppose that the mean of  $w$  is  $\mu_w$  and the standard deviation of  $w$  is  $\sigma_w$ . Then the center line, the upper control limit, and the lower control limit become

$$\begin{aligned} UCL &= \mu_w + L\sigma_w, \\ CL &= \mu_w, \\ LCL &= \mu_w - L\sigma_w, \end{aligned} \tag{1}$$

where  $L$  is the “distance” of the control limits from the center line, expressed in standard deviation units. This general theory of control charts was first proposed by Walter A. Shewhart, and control charts developed according to these principles are often called Shewhart control charts (Montgomery 2009). The control chart is a device for describing in a precise manner exactly what is meant by statistical control; as such, it may be used in a variety of ways. In many applications, it is used for on-line process surveillance. That is, sample data are collected and used to construct the control chart, and if the sample values of (say) fall within the control limits and do not exhibit any systematic pattern, we say that the process is in control at the level indicated by the chart (Burr 1976). What should be known here is that control charts are very important in the development of the process.

We can also use the control chart as an estimation tool. That is, from a control chart that exhibits statistical control, we may estimate certain process parameters, such as the mean, standard deviation, fraction nonconforming or fallout, and so forth. Montgomery (2009) noted that these estimates may then be used to determine the capability of the process to produce acceptable products. Such process-capability studies have considerable impact on many management decision problems that occur over the product cycle, including make or buy decisions, plant and process improvements that reduce process variability, and contractual agreements with customers or vendors regarding product quality.

## 2.2 Hotelling's $T^2$ Quality Control Chart

The multivariate quality control chart, developed by Harold Hotelling (1947), is based on  $T^2$  values with the assumption that the distribution of random variables is normal. This chart is often called the multivariate Shewhart control chart due to its resemblance to Shewhart control chart in many sources. It is also called a chi-square chart when it has a chi-square distribution with  $p$ -value degrees of freedom (Montgomery 2009; Eygü and Özçomak 2017). Hotelling  $T^2$  control graphs are one

of the multivariate quality control methods and are used to address more than one variable simultaneously and depict the variables on the same graph (Eygü 2015).

The simultaneous use of univariate Shewhart charts without taking the correlation between the variables into consideration may lead to a scheme where the Type I error is unknown. The corresponding multivariate SPC is more powerful in detecting a given shift. Consider  $p$  correlated characteristics as being measured simultaneously and these characteristics as being measured simultaneously and these characteristics follow a multivariate normal distribution with mean vector  $\mu'_0 = (\mu_{0,1}, \mu_{0,2}, \dots, \mu_{0,p})$ , and covariance matrix  $\Sigma_0$  when the process is in control, where  $\mu_{0,j}$  is the mean for the  $j$ th characteristic and  $\Sigma_0$  is a  $p \times p$  matrix consisting of the variances and covariance of the  $p$  characteristics. When an  $i$ th sample of size  $n$  is taken, we have  $n$  values of each characteristic and it is possible to calculate the  $\bar{X}_i$  vector, which represents the  $i$ th sample mean vector for the  $p$  characteristics. The charting statistic

$$T_i^2 = n(\bar{X}_i - \mu_0)' \Sigma^{-1} (\bar{X}_i - \mu_0) \tag{2}$$

is called Hotelling's  $T^2$  statistic that, when the process is in control, is distributed as a chi-square variate with  $p$  degrees of freedom ( $T_i^2 \approx X_p^2$ ). If  $T_i^2 > \chi_{p,\alpha=0.05}^2$  is in question, the process is considered to be out-of-control (Aparisi 2007).

Let us assume that  $\mu_x$  and  $\Sigma$  parameters are known to be in normal multivariate distribution, and suppose that  $\mu_x$  is a  $Y$  observation vector obtained from a normal multivariate distribution that has a different mean vector but the same covariance matrix. This observation vector is demonstrated by

$$T_Y^2 = (Y - \mu_x)' \Sigma^{-1} (Y - \mu_x). \tag{3}$$

The statistic here cannot be defined with  $\chi^2$ , the center of which is  $\mu_x$  and  $\mu_y$ . Thus, multivariate normal distribution defined with  $(Y - \mu_x)$  vector has a mean value that is different than 0. However, the center of the  $(Y - \mu_x)$  normal vector can be determined considering the mean  $\mu_x$  and  $\mu_y$  vectors and in parallel to the  $\bar{X}_1$  and  $\bar{X}_2$  axes. If we keep

$$(Y - \mu_x) = (Y - \mu_x + \mu_y - \mu_y) = [(Y - \mu_y) + \delta] \tag{4}$$

in mind,  $\delta = (\mu_y - \mu_x)$  indicates the mean deviation figure. With this result, the distribution of  $T_y^2$  is given by

$$T_y^2 = [(Y - \mu_y) + \delta]' \Sigma^{-1} [(Y - \mu_y) + \delta] \sim \chi_{(p,\lambda)}^2, \tag{5}$$

where  $\chi_{(p,\lambda)}^2$  is a non-central chi-square distribution with  $p$  degrees of freedom. A major difference between this distribution and the central chi-square is the additional parameter  $\lambda$ , labeled the non-centrality parameter (Mason and Young 2002). This parameter is demonstrated as  $\lambda = nd^2$ . When the in-control mean vector  $\mu_0$  shifts to

$\mu_1 = \mu_0 + \delta$  ( $\delta \neq 0$ ), the magnitude of this shift is often expressed by

$$d = \sqrt{(\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)} = \sqrt{\delta' \Sigma^{-1} \delta} \quad (6)$$

the Mahalanobis distance and  $\mu$  is the  $p$  characteristics mean vector. The subgroup statistic  $T_i^2$  then follows a non-central chi-square distribution with  $p$  degrees of freedom and non-centrality parameter  $\lambda = nd^2$ , that is  $T_i^2 \sim \chi_p^2(\lambda)$  (Rakitzis and Antzoulakos 2011). The average run length value of Hotelling's  $T^2$  control chart is dependent upon the mean  $\mu$  vector and covariance matrix,  $\Sigma$ , thanks to the non-central parameter demonstrated by  $d$ . In this equation,  $\mu_0$  indicates the target mean vector. It is possible to consider average run length as a function of  $d$ . We denote in-control state with  $d = 0$ ; hence, we have  $\alpha = P(T^2 > UCL | d = 0)$ . When the process is out of control with shift  $d \neq 0$ , we have  $\beta = P(T^2 < UCL | d \neq 0)$  (Faraz and Moghadam 2009). Faraz and Parsian (2006) suggested that schemes result in more rapid detection of lack of control and hence reduce the costs associated with nonconforming products. They showed that adding an additional warning line improves the performance of  $T^2$  control chart using variable sample size and variable sampling interval scheme (Eygü and Özçomak 2017).

### 2.3 Average Run Length (ARL)

ARL of the chart is a common measure of how well a chart performs in detecting an out-of-control process. Eygü and Özçomak (2017) states that ARL is commonly used to compare quality control charts. The run length is the number of the sampling stage at which the chart first signals. When begins to contemplate alternative schemes are considered to broadcast out-of-control signals in research based on process monitoring data, the need arises to quickly determine what a given scheme might be expected to do. The most effective means known for making this kind of prediction is the ARL notion. The first signal period of the process is expressed as  $T$  and  $T$  is called the run length for the scheme. The probability distribution of  $T$  is called the run length distribution and the mean or average value of this distribution is called the Average Run Length (ARL) for the process-monitoring scheme. That is ARL is defined as

$$ARL = \mu_T. \quad (7)$$

When one is setting up a process monitoring scheme, it is desirable that its procedure is a large ARL when the process is stable at standard values for process parameters and small ARLs under other conditions. Evaluating ARLs is usually not essential. However, there is a case where an explicit formula for ARLs is possible and where we can show the meaning and usefulness of the ARL concept in basic terms. Vardeman and Jobe (1999) stated that it is the situation where

- the process-monitoring scheme employees only the single alarm rule “signal the first time that a point  $Q$  plots outside control limits” and
- it is sensible to think of the process as physically stable (though perhaps not at standard values for process parameters).

Under the second condition, the values  $Q_1, Q_2, Q_3, \dots$  can be modeled as random draws from a fixed distribution, and the notation

$$p = P[Q_1 \text{ plots outside control limits}]$$

will prove useful. In this simplest of cases, Mason and Young (2002) explained that ARL for a control procedure is defined as

$$ARL = \frac{1}{p}, \tag{8}$$

where  $p$  represents the probability of being outside the control region. For a process that is in control, this probability is equal to  $\alpha$ , the probability of a Type I error. The ARL has a number of uses in both univariate and multivariate control procedures. Javaheri and ve Houshmand (2001) determined different covariance structures using  $p$ -dimensional multivariate data. Various amount of shift in the mean vector is induced, and the resulting ARL is computed. They evaluated the effectiveness of ARL by considering five different methods including Hotelling’s  $T^2$ , Shewhart Control Charts, Discriminant Analysis, Decomposition Method, and Multivariate Ridge Residual Chart.

Two cases are presented for ARL. The first is that the process mean is at the targeted values which is called controlled ARL and indicated as  $ARL_0$ . The second case is when the process mean deviates—called out-of-control ARL and indicated as  $ARL_1$ . If the process mean is at the targeted values, the signal indicated by the control chart is false. Thus, the expected ARL value should be large in this case. When the process mean deviates, the signal indicated by the control chart is correct; thus, the expected ARL value should be small (Cox 2001).

ARL can also be used to calculate the number of mean expected observations before an out-of-control signal is present in the process, in other words when the process is in control. Thus, the ARL value can be obtained using the following:

$$ARL_0 = \frac{1}{\alpha}. \tag{9}$$

When a shift is present in the process, another use of ARL is to calculate the number of observations of the shift before the shift itself is detected. The probability of detecting any shift, and possible deviation for the  $\beta$  value is equal to  $(1-\beta)$ . The  $\beta$  value represents Type II error probability. If a shift is present, this probability deviations can be determined using standard statistical equals. The ARL for detecting the shift is given by

$$ARL_1 = \frac{1}{1 - \beta}. \quad (10)$$

The probability  $(1-\beta)$  represents the power of the test of a statistical hypothesis that the mean has shifted. This result produces another major use of the ARL, which consists of comparing one control procedure to another. This is done by comparing the ARLs of the two procedures for a given process shift (Mason and Young 2002). As will be seen, the probability  $\beta$  is a function of the distributional parameters  $\mu$ ,  $\mu_0$ , and  $\Sigma_0$  of the distribution of the vector of quality measurements,  $X$ , only through the value  $d$ , where  $d = \sqrt{(\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)}$  with  $\mu$  the out-of-control value of the mean vector provided  $\Sigma = \Sigma_0$  (Champ and Aparisi 2008: 155, Eygü and Özçomak 2017: 3).

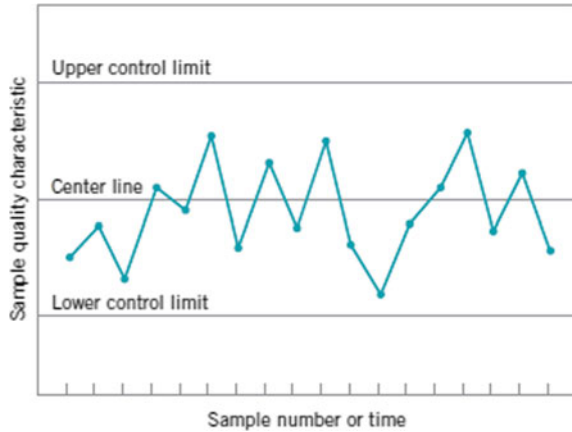
### 3 Materials and Methods

#### 3.1 Data Description

This study utilized sugar production data in the Erzurum provinces in Turkey. The corresponding data were obtained from the Erzurum sugar factory and were extracted from the sugar production reports. The statistical package JMP from SAS was used to analyses and graph all the charts given in this study. In each report, relevant information about the sugar quality was categorized under five variables affecting sugar quality. After the underlying data were classified into a convenient, computer excel form, among 200 sugar data were analyze, which were selected according to the method of Simple Random Sampling (SRS) and Ranked Set Sampling (RSS). In statistical surveys, if the measurements of sampling units according to the variables under consideration is expensive in all sense, and if it is possible to rank sampling units according to the same variable by means of method which is not expensive at all, in those cases, RSS is more efficient than SRS as a sampling method in the sense of estimation the population mean. Further this increase is the recognition by statisticians of the need for more cost-effective sampling procedures, such as those that use a prior knowledge or can otherwise provide the needed information with a significant reduction in cost over the more traditional sample random sampling approaches. The goal of RSS is to collect observations from a population that are more likely to span the full range of values in the population than same number of observations obtained simple random sampling. Firstly, according to SRS and RSS the samples were selected with excel program. In order to create RSS example, a ranked set sample design with set size  $m = 5$  and number of sampling repeat  $r = 10$ . Although 50 sample units have been selected from the population, only the 50 units SKÖ sample are actually included in the final sample for quantitative analysis. If we quantified the sample number of sample units,  $mr$ , by a simple random sample, then we have no control over which units enter the sample. Then this sample was



**Fig. 1** A typical control chart



analyzed by Hotelling  $T^2$  and multivariate CUSUM control chart. As seen in Fig. 1. Samples are randomly selected for 5 sets, where each set contains 5 sample units, and then repeat this procedure for 10 times.

Figure 2, where each row denotes a judgment-ordered sample within a repeat, and the units selected for quantitative analysis are repeated. This situation was repeated cycles  $m = 5$  and  $r = 20$  for Hotelling  $T^2$ . For the SKÖ, it was repeated cycles  $m = 5$  and  $r = 10$ . Obtaining a sample in this manner maintains the unbiasedness of SRS; however, by incorporating outside information about the sample units, we are able to contribute a structure to the sample that increases its representativeness of the true underlying population.

### 3.2 Quality Control Charts Using RSS and SRS

Let  $X_{ij}$  for  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, r$  denote the  $i$ th unit the  $j$ th SRS of size  $n$  and  $X_{ij} \sim N(\mu, \sigma^2)$ . If the population mean  $\mu$  and variance  $\sigma^2$  are known, the Shewhart control chart for

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}; \quad j = 1, 2, \dots, r \tag{11}$$

is given by

$$\begin{aligned} UCL &= \mu + 3 \frac{\sigma}{\sqrt{n}}, \\ CL &= \mu, \\ LCL &= \mu - 3 \frac{\sigma}{\sqrt{n}}, \end{aligned} \tag{12}$$

Round	Rank				
	1	2	3	4	5
<i>r</i> = 1	∇	•	•	•	•
	•	∇	•	•	•
	•	•	∇	•	•
	•	•	•	∇	•
	•	•	•	•	∇
<i>r</i> = 2	∇	•	•	•	•
	•	∇	•	•	•
	•	•	∇	•	•
	•	•	•	∇	•
	•	•	•	•	∇
<i>r</i> = 10	∇	•	•	•	•
	•	∇	•	•	•
	•	•	∇	•	•
	•	•	•	∇	•
	•	•	•	•	∇

Fig. 2 A ranked set sample design with set size  $m = 5$  and  $r = 10$

where UCL, CL, and LCL denote the upper control limit, central limit, and lower control limit respectively. The sample means  $\bar{X}_j$   $j = 1, 2, \dots, r$  can be plotted in the above charts. For this chart average run length is equal to  $1/\alpha$ , where  $\alpha$  is the probability of type I error if the process is under control. But if process mean shift then  $ARL = 1/(1 - \beta)$ , where  $\beta$  is the probability of type II error (Montgomery 2009; Muttlak and Al-Sabah 2003).

The RSS scheme consists of drawing  $n$  random samples, each of size  $n$  from target population, and ranking the units within each set with respect to a variable of interest. The RSS, mean of the  $j$ th cycle can be plotted on the control on the control chart based on RSS data suggested by Salazar and Sinha 1997 for a discussion of the impact,

$$\begin{aligned}
 UCL &= \mu + 3\sigma_{\bar{X}_{RSS}}, \\
 CL &= \mu, \\
 LCL &= \mu - 3\sigma_{\bar{X}_{RSS}},
 \end{aligned}
 \tag{13}$$

where  $\sigma_{\bar{X}_{RSS}} = \sqrt{(1/n^2) \sum_{i=1}^n \sigma_{(i:n)}^2}$  and is calculated using the known results from the tables of order statistics for standard normal distribution (Harter and Balakrishnan 1996).

### 3.3 Computation for Hotelling $T^2$

Graphical procedure that is helpful in assessing if a set of data represents a reference distribution is a control chart. Thus, this technique can be used in assessing the sampling distribution of the  $T^2$  statistic (Gnanadesikan 1977; Sharma 1995). To illustrate the above procedure when using individual observations, consider the 20 observations presented in Table 1. The  $T^2$  values of the 20 observations are computed using and are presented in Table 1. Using an  $\alpha = 0.01$ , the control limit is computed using the formula in (1).

Observations in RSS 5 are detected as an outlier since its  $T^2$  value of 14.75 exceeds the UCL. Graphical approach to outlier detection based on the  $T^2$  statistics is to examine a control chart of the appropriate  $T^2$  values. For SRS and RSS, the  $T^2$  values of the preliminary data set are presented in the  $T^2$  control chart given in Figs. 3 and 4.

To compare our suggested control charts with the classic charts, we collected data using the SRS method for sample. We construct the quality control charts using the classical method to find the upper limits (see Montgomery 2009). Figure 3 shows the  $T^2$  control charts that no points are outside the upper control limit, and Fig. 4 shows the  $T^2$  control charts that one point is outside the upper control limit. In contrast to

**Table 1**  $T^2$  values for sampling using SRS and RSS ( $m = 5, r = 20$ )

Sample no	SRS			RSS			
	$T^2_{0.01;10.22}$			$T^2_{0.01;11.31}$			
1	7.75	11	0.83	1	1.82	11	6.87
2	3.07	12	5.33	2	5.49	12	4.52
3	10.11	13	2.83	3	4.29	13	1.74
4	2.51	14	3.56	4	2.61	14	1.49
5	8.25	15	5.07	5	14.75*	15	11.04
6	4.87	16	6.06	6	9.22	16	3.34
7	3.56	17	3.27	7	0.35	17	6.61
8	4.38	18	6.11	8	2.74	18	2.30
9	3.15	19	3.15	9	2.19	19	3.80
10	5.62	20	5.89	10	1.24	20	8.54

\*Indicates  $T^2$  value is significant at 0.01 level

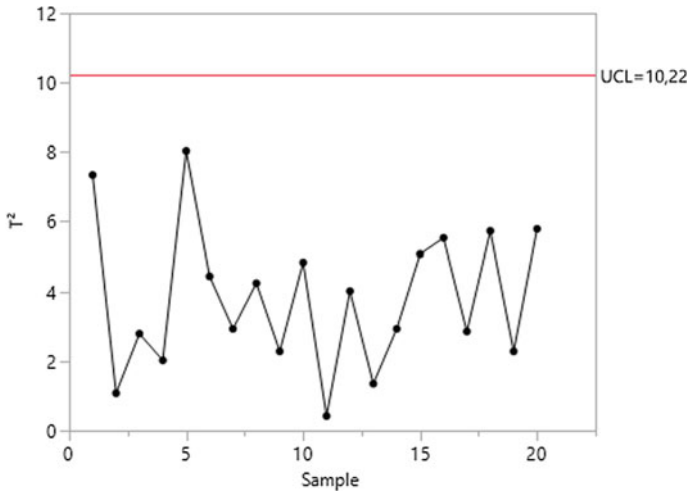


Fig. 3 Hotelling  $T^2$  control chart using SRS

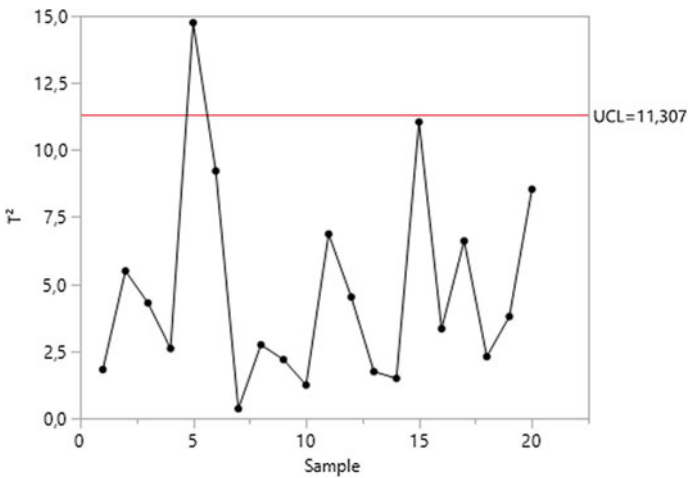


Fig. 4 Hotelling  $T^2$  control chart using RSS

the control chart on the individual measurements in Fig. 3, we would conclude that this process is in a reasonable state of statistical control. As a result of analysis seems that the RSS method is doing a better job in estimating the  $T^2$  than the SRS.

### 3.4 Computation for CUSUM

The primary Shewhart control chart rule, look for trouble if a point falls outside the control limits, is a statistically independent decision rule. That is, the decision is based on appoint plot of information taken from one and only one sample or subgroup. The CUSUM technique may be illustrated in its most favorable light by an example in which samples are taken for a period of time from a stable universe that is known to be normality distributed. This period is followed by one which the centering of the universe has changed to a different level and there has been no change in the universe dispersion (Grant and Leavenworth 1996). Statistical control chart gives us some very definite information about these matters. It tells us that in the control chart the average of  $\bar{X}$  values will be the same as  $\mu$ , the average of the universe. CUSUM charts of the type described above have been in use for more than years, particularly in the study of chemical processes. Montgomery (2009) stated that CUSUM is designed by choosing values for the reference value  $K$  and the decision interval  $H$ . It is usually recommended that these parameters be selected to provide good average run length performance. There have been many analytical studies of CUSUM ARL performance. Based on these studies, we may give some general recommendations for selecting  $H$  and  $K$ . Define  $H = h\sigma$  and  $K = k\sigma$ , where  $\sigma$  is the standard deviation of the sample variable used in forming the CUSUM. For example, using  $h = 4$  or  $h = 5$  and  $k = 0.5$  will generally provide a CUSUM that has good ARL properties against a shift of about  $1\sigma$  in the process mean. Pignatiello and Runger (1990) showed that the Woodall and Ncube (1985) multiple CUSUM chart does not have good average run length (ARL) properties when the process mean shifts along several characteristics simultaneously. To lessen the sensitivity of the multiple univariate CUSUM chart to directions, they recommended using the univariate CUSUM charts aimed at several uniformly elected directions. But, at the same time, they found that the resulting control chart is hard to manage when there are three or more characteristics. Hawkins (1993) indicated that under some circumstances separate controls on the regression—adjusted variables by the CUSUM charts can both improve the speed of detection and make the chart signal more easily interpretable (Hamed et al. 2016).

The CUSUM chart directly incorporates all the information in the sequence of sample values by plotting the cumulative sums of the deviations of the sample values from a target value. Supposed the samples of size  $n \geq 1$  are collected, and  $\bar{X}_j$  is the average of the  $j$ -th sample. Then if  $\mu_0$  is the target for the process mean, the cumulative sum control chart is formed by plotting the quantity

$$C_i = \sum_{j=1}^i (\bar{x}_j - \mu_0) \quad (14)$$

against the sample number  $i$ .  $C_i$  is called the cumulative sum up to and including the  $i$ th sample.

Because they combine information from several samples, cumulative sum charts are more effective than Shewhart charts for detecting small process shifts. Furthermore, they are particularly effective with samples of size  $n = 1$ . This makes the cumulative sum control chart a good candidate for use in the chemical and process industries where rational subgroups are frequently of size 1, and in discrete parts manufacturing with automatic measurement of each part and on-line process monitoring directly at the work center (Montgomery 2009).

Montgomery (2009) showed that a tabular CUSUM may be constructed for monitoring the mean of a process. The tabular CUSUM works by accumulating derivations from  $\mu_0$  that are above target with one statistic  $C^+$  and accumulating derivations from  $\mu_0$  that are below target with another statistic  $C^-$ . The statistics  $C^+$  and  $C^-$  are called one-sided upper and lower CUSUMS, respectively. They are computed as follows: The tabular CUSUM,

$$C_i^+ = \max[0, x_i - (\mu_0 + K) + C_{i-1}^+], \quad (15)$$

$$C_i^- = \max[0, (\mu_0 - K) + C_{i-1}^-], \quad (16)$$

where the starting values are  $C_i^+ = C_i^- = 0$ .

In Eqs. (12) and (13),  $K$  is usually called the reference value (or the allowance, or the slack value), and it is often chosen about halfway between the target  $\mu_0$  and the out of control value of the mean  $\mu_1$  that we are interested in detecting quickly. Thus, if the shift is expressed in standard deviation units as

$$\mu_1 = \mu_0 + \delta\sigma \text{ (or } \delta = |\mu_1 - \mu_0|/\sigma),$$

then  $K$  is one-half the magnitude of the shift or

$$K = \frac{\delta}{2}\sigma = \frac{|\mu_1 - \mu_0|}{2}. \quad (17)$$

Note that  $C_i^+$  and  $C_i^-$  accumulate deviations from the target value  $\mu_0$  that are greater than  $K$ , with both quantities reset to zero on becoming negative. If either  $C_i^+$  or  $C_i^-$  exceed the decision interval  $H$ , the process is considered to be out of control.

In this section, we used statistics program to compare the CUSUM using the usual simple random sampling (SRS) and ranked set sampling (RSS) data to the newly developed CUSUM control charts. To be able to compare the values of the average run length (ARL) using RSS with existing ARL using SRS, we used in program the values of  $\delta$ ,  $k$ , and  $h$ . Hawkins and Olwell (1998) used in our simulation the same values of  $\delta$ ,  $k$ , and  $h$ . Harter and Balakrishnan (1996) showed that Monte Carlo studies conducted for several of these populations in order to check the applicability of the asymptotic theory to samples of small or moderate size and to compare the maximum-likelihood estimators with other estimators. ARL is generated from the best chart, and results obtained are displayed at the bottom of Tables 2 and 3.

**Table 2** ARL values for the Shewhart-CUSUM scheme using SRS

$\delta$	$k = 0.25, h = 7.50$	$k = 0.50, h = 5$	$k = 0.75, h = 3$
0	282.96	465.42	221.40
0.5	26.78	38.01	39.31
1.0	10.73	10.38	9.68
1.5	6.71	5.75	4.73
2.0	4.93	4.01	3.12
2.5	3.93	3.11	2.36
3.0	3.30	2.57	1.93

**Table 3** ARL values for the Shewhart-CUSUM scheme using RSS ( $m = 5, r = 10$ )

$\delta$	$k = 0.25, h = 7.50$	$k = 0.50, h = 5$	$k = 0.75, h = 3$
0	278.49	420.57	205.19
0.5	26.67	36.77	37.92
1.0	10.53	10.18	9.50
1.5	6.59	5.65	4.66
2.0	4.82	3.94	3.08
2.5	3.86	3.06	2.33
3.0	3.24	2.53	1.91

Tables 2 and 3 show that CUSUM control charts based on RSS gives better ARL performance as compared to their corresponding control charts for mean using SRS when a process is out of control,  $\delta > 0$ . Observe that as shift  $\delta$  increases in RSS based control charts, the ARL values decreases. Lee (2013) showed that the ARL values for the RSS decrease much faster than SRS if  $\delta$  increases. In real life we do not know  $\mu$  and  $\sigma_{\bar{x}_{RSS}}$ , we need to estimate them using RSS data. A recent study developed that control charts dominate the classical charts. If the process starts to get out of control by reducing the number of ARL substantially. But number of false alarms is not reduced by the same amount if the process is under control.

## 4 Results

Hotelling's  $T^2$  control chart has the advantage of its simplicity, but it is slow in detecting small process shift. The latest developments in variable sample sizes for univariate control charts are applied in this study to define an adaptive sample sizes  $T^2$  control chart. As occurs in the univariate case the ARL improvements are very important particularly for small process shift. The user has been provided with tables of these schemes that easily allow an optimal plan. Ranked set sampling has been demonstrated to be an efficient sampling method. Following the most recent literature (Faraz and Moghadam 2009) the proposed schemes result in more rapid detection

of lack of control and hence, the costs associated with nonconforming products. The results show that the sampling plan to be applied is a function of the magnitude of the process shift. Following the same procedure is consistent with results from Muttlak and Al-Sabah (2003) develop different quality control charts for the sample mean using ranked set sampling (RSS). These charts were compared to the usual control charts on RSS, or one of its modifications is shown to have smaller ARL than classical chart when there is a sustained shift in the process mean. The results show that the sampling plan to be applied is a function of the magnitude of the process shift. For small process shifts, we should employ large sample sizes infrequently, and for large process shifts, a small sample size should be taken very frequently. This sampling method will help detect the defective products in time and minimize the cost and loss for establishments.

## References

- Aparisi F (2007) Sampling plans for the multivariate  $T^2$  control chart. *Qual Eng* 10(1):141–147
- Burr IW (1976) *Statistical quality control methods*, 16th ed. CRC Press, Boca Raton
- Champ CW, Aparisi F (2008) Double sampling Hotelling's  $T^2$  charts. *Qual Reliab Eng Int* 24(2):153–166
- Cox MAA (2001) Towards the implementation of a universal control chart and estimation of its average run length using a spreadsheet: an artificial neural network is employed to model the parameters in a special case. *J Appl Stat* 28(3):353–364
- Eygü H (2015) Application of multivariate statistical process control in cement industry. *J Bus Econ Polit Sci* 4(8):67–82
- Eygü H, Özçomak MS (2017) Multivariate statistical quality control based on ranked set sampling. *Asian Soc Sci* 14(1):2–5
- Faraz A, Moghadam MB (2009) Hotellin's  $T^2$  control chart with two adaptive sample sizes. *Qual Quant* 43(6):903–904
- Faraz A, Parsian A (2006) Hotelling's  $T^2$  control chart with double warning lines. *Stat Pap* 47:569–593
- Gnanadesikan R (1977) *Methods for statistical data analysis of multivariate observations*. Wiley, New York
- Grant EL, Leavenworth RS (1996) *Statistical quality control*, 7th edn. McGrawHill, Columbus, OH
- Hamed MS, Mansour MM, Elrazik E (2016) MCUSUM control chart procedure: monitoring the process mean with application. *J Stat Adv Theory Appl* 16(1):105–132
- Harter HL, Balakrishnan N (1996) *CRC handbook of tables for the use of order statistics in estimation*. CRC Press, Boca Raton
- Hawkins DM (1993) Regression adjustment for variables in multivariate quality control. *J Qual Technol* 25(1993):170–182
- Hawkins DM, Olwell DH (1998) *Cumulative sum charts and charting improvement*. Springer, New York
- Javaheri A, ve Houshmand AA (2001) Average run length comparison of multivariate control charts. *J Stat Comput Simul* 69:125–140
- Lee MH (2013, April) The three statistical control charts using ranked set sampling. In: 2013 5th international conference on modeling, simulation and applied optimization (ICMSAO), IEEE, pp 1–6
- MacGregor JF, Kourti T (1995) Statistical process control of multivariate processes. *Control Eng Pract* 3(3):403–414



- Mason RL, Young, JC (2002) Multivariate statistical process control with industrial applications, ASA-SIAM Series on Statistics and Applied Probability
- Montgomery CD (2009) Introduction to statistical quality control, 6th ed. Arizona State University, Wiley, United States of America, pp 180–200
- Muttlak H, Al-Sabah W (2003) Statistical quality control based on ranked set sampling. *J Appl Stat* 30(9):1055–1078
- Noorossana R, Vaghefi SJM (2006) Effect of autocorrelation on performance of the MCUSUM control chart. *Qual Reliab Eng Int* 22(2):191–197
- Pignatiello JJ, Runger GC (1990) Comparison of multivariate CUSUM charts. *J Qual Technol* 22:173–186
- Rakitzis CA, Antzoulakos LD (2011) Chi-square control charts with runs rules. *Methodol Comput Appl Probab* 13(4):657–669
- Sharma S (1995) Applied multivariate techniques. Wiley, New York
- Shewhart WA (1931) Economic control of quality of manufactured products. In: Van Nostrand D, Princeton NJ (eds) Reprinted by the American society for quality control. Mil-waukee, Wis
- Vardeman SB, Jobe JM (1999) Statistical quality assurance methods for engineers. Wiley, New York
- Woodall WH, Ncube MM (1985) Multivariate CUSUM quality-control procedures. *Technometrics* 27:285–292

# A Robust Regression Method Based on Pearson Type VI Distribution



Yasin Büyükkör and A. Kemal Şehirlioğlu

**Abstract** In classical regression analysis, the distribution of the error is assumed to be Gaussian, and Least Squares (LS) estimation method is used for parameter estimation. In practice, even if the distribution of errors is assumed to be Gaussian, residuals are not generally Gaussian. If the data set contains outlier (s) or there are observations that are suspected to be outlier, normality assumption is violated, and parameter estimates will be biased. Many statisticians used robust method, such as the M-Estimation Method, which is a generalized version of the Maximum Likelihood (ML) Estimation method, for parameter estimation when such problems occurred. However, if the data set has skewness and excess kurtosis, traditional M-Estimators cannot achieve a good solution. In this study, using the relationship between Pearson Differential Equation (PDE) and Influence Function (IF), M-Estimation method is proposed for data sets that follow Pearson Type VI (PVI) distribution. The advantage of this method takes into account the skewness and kurtosis values of the data set and generates dynamic solutions. Objective, influence, weight functions and tail properties of the PVI distribution are obtained by using the Probability Density Function (pdf) of the PVI distribution. For the regression parameter estimates, Iteratively Re-Weighted Least Squares Estimation Method (IRWLS) is used. In many simulation studies with different scenarios and applications with real data, if the data have skewness and excess kurtosis, the proposed method has achieved better results than other M-Estimation methods in terms of Total Absolute Deviation (TAB) and Mean Square Error (MSE).

**Keywords** M-Estimation method · Robust regression · Pearson type VI distribution · Influence function · Iteratively re-weighted least squares method

---

Y. Büyükkör (✉)

Faculty of Economics and Administrative Sciences, Department of Econometrics, Karamanoğlu Mehmetbey University, 70200 Karaman, Turkey

e-mail: [yasinbuyukkor@kmu.edu.tr](mailto:yasinbuyukkor@kmu.edu.tr)

A. K. Şehirlioğlu

Faculty of Economics and Administrative Sciences, Department of Econometrics, Dokuz Eylül University, 35210 İzmir, Turkey

e-mail: [kemal.sehirli@deu.edu.tr](mailto:kemal.sehirli@deu.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

117

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_8](https://doi.org/10.1007/978-3-030-85254-2_8)

## 1 Introduction

In the history of statistics, many researchers have analyzed the data assuming Gaussian distribution. Thus, anomalies in the data (heavy-long tail, excess kurtosis, skewness, outlier, etc.) are often ignored by researchers. Such anomalies can be caused by many reasons. The main reasons are measurement and recording mistakes or mixing of two or more populations. However, in the data set, an observation (s) belonging to the data set can act like an abnormal observation. Researchers have difficulty in analyzing the data set in the presence of such anomalies and they use analysis methods ‘robust’ to anomalies in order to overcome such situations. Robust statistics is concerned with deviations from the assumed model and the construction of reliable and sufficiently efficient statistical procedures when these deviations occur. The term ‘Robust’ was first used by Box (1953). Tukey (1960) observed that even small perturbations from the assumed model cause optimal procedures to rapidly lose their effectiveness, and Tukey (1962) has led the robust methods used today. Huber (1964) developed the M-estimator, a flexible and broad class of estimators, which has an important place in the development of robust statistical methods. Hampel (1968, 1974) introduced the Influence function, which is one of the most important tools in measuring the stability of a statistical procedure and has played an important role in the development of new robust methods. M-Estimators are frequently used in Theoretical and Applied Statistics, Econometrics and Biostatistics.

In the regression analysis, the anomalies in the data while estimating the parameters can cause to lose the effectiveness of the LS estimation method. In the presence of such data, parameter estimates made with OLS are will be biased (Hampel 1968). In robust statistics, traditional M-estimation methods do not consider the skewness and kurtosis parameters of the data, the PDE contains these values. Thus, the distribution of the data set can be determined uniquely, and the error is minimized while estimating the regression parameter.

The aim of this study is to construct a new method based on PVI that can be used instead of conventional M-estimators when data have anomalies. While traditional M-estimators usually achieve a good solution for symmetric and heavy long-tailed data, they lose effectiveness when anomalies arise. Therefore, in this study, regression parameter estimates will be estimated by using the weight function of the PVI, which contains the asymmetry, kurtosis and heavy long tail occurring in the data set. Based on the similarity between PDE and IF, Objective Function, Influence Function and Weight Function will be obtained by using the pdf of PVI.

This paper is organized as follows: Sect. 2 outlines the literature review about Robust Regression and Pearson Distribution System (PDS), Sects. 3 and 4 outlines the theoretical framework of Robust Regression and PDS, Sect. 5 presents the relationship between PDE and IF, and also for the proposed method, which based on PVI, obtained Objective, IF and Weight Function, Sect. 6 presents two real-world examples and simulation study with different scenarios, also discussion on obtained results for the proposed method. In the last section, we discuss the advantages and

disadvantages of the proposed method. This paper also has an Appendix section, which contains proof of the tail properties of PVI.

## 2 Literature Review

Robust regression analysis has been studied frequently in the literature, especially after the 1960s. After Tukey, Huber and Hampel, many researchers have been interested in robust regression analysis. To summarize briefly, Harvey (1977) suggested using the minimum absolute deviation estimator as an initial solution in the robust regression procedure. M-estimators based on the median developed by Hinich and Talwar (1975) and Andrews (1974) were also used as the initial solution. Hogg (1979) discussed the robust statistical procedures used to reduce the effects of outliers in the data set. He examined the estimation processes of regression parameters and focused on the IRWLS method, which is the method used to estimate regression parameters, and discussed the asymptotic variance formula. He discussed the data set, which is reported by Andrews (1974) and analyzed by Wood and Gorman (1971). In addition, he used M-Estimator for analysis of the data sets which ‘Half-life of Plutonium-241’ by Zeigler and Ferris (1973) and ‘Splines’ by Lenth (1977), and ‘Automated data reduction’ by Agee and Turner (1978). Wu (1985) discussed commonly used M-estimators for scale and regression parameters. He compared the Bell/OLS M-Estimators developed by Bell (1980) and the high breakdown point Bell/RM M-estimators developed by Siegel (1982) using several real data sets. He discussed the similarities between Tukey Bisquare M-estimator and the Bell/OLS M-estimators. Croux and Reusseeuw (1992) developed two robust scale estimates,  $S_n$  and  $Q_n$ . They focused on breakdown points and computational algorithms for the developed estimators. They compared these estimators according to calculation time. They also used these scale estimates while estimating regression parameters. Cantoni and Ronchetti (2006) have developed a new robust method to be used in skewed and heavy-tailed data. They proved that when there are deviations from the assumed model, the method they developed is more efficient than traditional methods. They demonstrated the efficiency of the method they developed by using “medical back problems” data obtained from 100 patients in a hospital in Switzerland and many simulation studies.

Allende et al. (2006) proposed an M-estimation method with an asymmetric influence function based on the  $G_A^0$  distribution. They used the developed method to process images obtained from satellite (GPS). Mohebbi et al. (2007) examined the robust regression methods that are an alternative to LS. They compared Least Absolute Deviation (LAD), Huber and nonparametric regression methods using skewed data sets. They used MSE and TAB as comparison criteria. Chen (2013) suggested using the distributed (clustered) IRWLS estimation method, when the data set is very large. Rasheed et al. (2014) used IRWLS to estimate regression parameters in the presence of outlier or heteroscedasticity in the data set. They also compared M-Estimator, LS and Least Trimmed Squares (LTS) methods using different data

sets. Khalil et al. (2016) proposed a redescending M-estimator. He compared this estimator with the Hampel, Andrews, Tukey and Qadir M-estimators. In addition to many simulation studies, they compared the methods using the data set of international telephone calls from Belgium (Rousseeuw and Leroy 1987) between 1950 and 1973. Sumarni et al. (2017) studied the location parameter of the distribution as robust using the T distribution, which has a longer tail than the normal distribution and obtained the Objective, Influence and Weight functions of the T-distribution. They obtained the asymptotic behavior and Asymptotic Relative Efficiency (ARE) for location parameter. They examined how ARE changes using different degrees of freedom. Yulita et al. (2018) compared the weight functions of Huber, Hampel, Tukey and Welsch using simple and multivariate regression analysis. They used many simulations and Human Development Index (HDI) data from India East Java Region for comparison. Considering the literature for PDS, Pearson Differential Equation, first introduced by Karl Pearson (1895), is a system that generates different probability distributions according to the different values of the parameters in the PDE. This system is called the Pearson Distribution Family (PDF) and includes 13 different distributions with 3 are main types and the 10 transition types. The main types of PDF:

- Pearson Type I Distribution (PI),
- Pearson Type IV Distribution (PIV),
- Pearson Type VI Distribution (PVI).

PI (Four Parameter Beta Distribution) is a limited distribution from both tails. The PIV, on the other hand, is a distribution whose roots are complex, but it is unlimited at both tails. The PVI distribution (Beta Distribution the Second Type) is a heavy long-tailed distribution (see Appendix.) limited in one tail (right or left). PVI contains F, Pareto, Beta and Gamma Distributions according to the values of the parameters of the distribution. In addition, due to the structure of its parameters, it can be used easily in many kurtosis and skewness values. Mainly used areas as follows:

- Loss Function (Balkema and Embrechts 2018),
- Examination of Brain Functions (Brascamp et al. 2004),
- Modeling in Epidemic Diseases (Tulupyev et al. 2013),
- Meteorology and Hydrology (Mielke and Johnson 1974),
- Financial Volatility (Moghaddam et al. 2019),
- Income Modeling (Ye et al. 2012),
- Processing of Radar Images (Salazar 2000) and
- Reliability Analysis (Kilany 2016).

### 3 Robust Regression

The development of robust methods has led to significant improvements in regression analysis as in all other statistical methods. Especially when the data contain outliers, it has become inevitable to use robust methods. It has been difficult for researchers

to define an observation in the data set as an outlier. Barnett and Lewis (1984) stated for the outliers as “inconsistent observations for the rest of the data set”. Judge et al. (1988) called the large values in regression residuals the outliers. According to Hampel et al. (1986) and Krasker et al. (1983), outliers are divided into two groups as gross errors and model errors. Gross errors are errors due to recording, writing, failure of measuring equipment, unit change or misinterpretation. Even a small number of gross errors in the data set cause a tremendous change in traditional LS estimators. In the presence of such situations, it is of great importance to use robust statistical methods. Model error may occur due to the structure of the statistical/econometric model, such as misinterpretation of a variable or removed variable, which is the great contribution of the model.

In Fig. 1, the observations within the black circle are vertical (y-direction) outliers. The values of these observations  $x_i$  are close to rest of the data. However, these values do not follow the linear relationship that most of the data have. The observations within the red circle are points of “good leverage”. They have the linear relationship that most of the data but have great  $x_i$  values. Contrary to its good name, they have a great influence on the LS estimators. Observations within the green circle are points of “bad leverage”. They have large  $x_i$  values and do not fit most of the data set. They have a tremendous influence on the LS estimators. They could be gross errors.

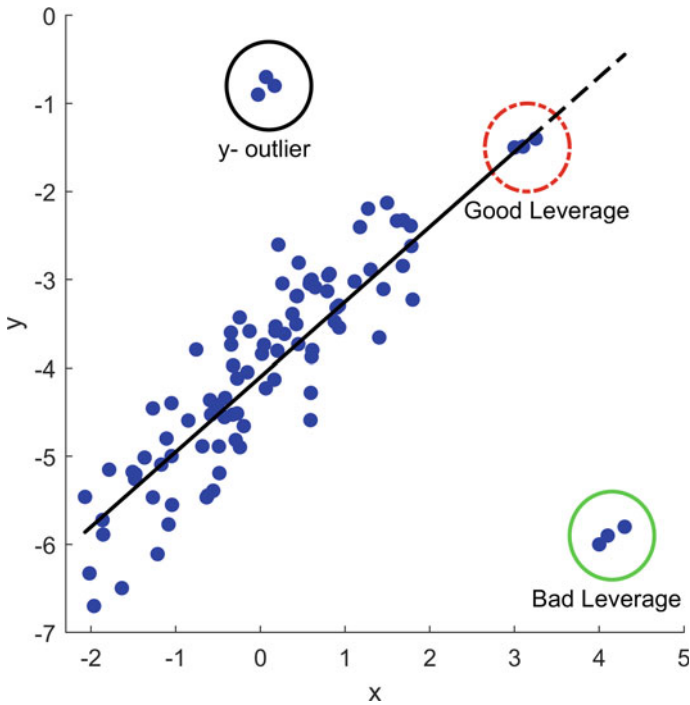


Fig. 1 Regression outliers

It is important to use robust methods in cases such as gross errors or model errors to minimize the effects of these errors. Features that robust regression estimators should have:

- If there are no outliers in the data and the distribution is normal, it should have a good performance as LS.
- When the first condition is not met, should have a better performance than the LS.
- Understanding the theory should be at least as easy as the LS method.
- It should be insensitive to trivial perturbations in the data.
- It should be easily calculated (Ryan 2008; Staudte and Sheather 2011).

### 3.1 Regression M-Estimator

If the distributions of the errors are heavy-tailed or there are outliers in residuals, parameter estimates made by LS will be biased (Hampel et al. 1986). Many researchers use robust methods to overcome such problems arise. One of the most popular robust methods is M-Estimators, which is based on ML proposed by Huber (1964) (Stuart 2011; Andersen 2008).

Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (1)$$

where  $\mathbf{y}$  is an  $n \times 1$  response vector,  $\boldsymbol{\theta}$  is an  $p \times 1$  unknown regression parameters,  $\mathbf{X}$  is an  $n \times 1$  explanatory variable matrix and  $(\mathbf{X}^T\mathbf{X})^{-1}$  is of full rank and  $\boldsymbol{\varepsilon}$  is an  $n \times 1$  error vector. In the classical LS method minimizing sum of squares:

$$\arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 \quad (2)$$

Differentiating Eq. (2) with respect to  $\boldsymbol{\theta}$  and system of  $p$  equations can be obtained:

$$\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta}) \mathbf{x}_{ij} = 0 \quad (3)$$

Solving Eq. (3) with respect to  $\boldsymbol{\theta}$ :

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (4)$$

In Robust Regression Analysis, we can maximize or minimize the different functions or distributions of errors instead of minimizing the sum of squares of errors:

$$\sum_{i=1}^n \rho(y_i - \mathbf{x}_i^T \boldsymbol{\theta}) = \min! \tag{5}$$

where  $\rho = -\ln f(x)$  and can be defined as Objective Function. (Susanti and Pratiwi 2014). Differentiating Eq. (5) with respect to  $\boldsymbol{\theta}$ :

$$\sum_{i=1}^n \psi(y_i - \mathbf{x}_i^T \boldsymbol{\theta}) x_{ij} = 0 \tag{6}$$

where  $\psi(\cdot)$  is Influence or Score Function. Solving Eq. (6) and obtaining  $i$ -th residuals is  $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\theta}}$ , one can rewrite the Objective and Influence Function as follows, respectively:

$$\min \sum_{i=1}^n \rho\left(\frac{e_i}{s}\right) \tag{7}$$

$$\sum_{i=1}^n \psi(r_i) x_{ij} = 0 \tag{8}$$

where  $r_i = e_i/s$  and  $s$  is the estimation of standard deviation ( $\sigma$ ) must be the use for scale equivariance. Even if there are many different  $s$  estimates, the Median Absolute Deviation (MAD), which is not affected by outliers, is the most widely used for scale estimation (Draper and Smith 2014). MAD can be written as:

$$s = MAD/0.6745 = \text{median}|e_i - \text{median}(e_i)|/0.6745 \tag{9}$$

where 0.6745 is correction constant for the data actually normal (Hogg 1979).

If Eq. (8) can be written as a weighted LS estimation problem:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \tag{10}$$

where  $w_i = \psi(r_i)/r_i$  and  $\mathbf{W} = \text{diag}\{w_i, i = 1, \dots, n\}$  is a  $n \times n$  weight matrix (Huber and Ronchetti 1981). The Weighted Least Squares method is usually used when solving Eq. (10). However, in order to calculate the weights, the first solution should be made with an appropriate method (usually LS) and the weights should be calculated with the weight function of the selected M-Estimators.

In the regression analysis, the distribution of errors is generally assumed to be Gaussian. If the distribution of the errors is actually normal, MLE and LS methods are the same. M-Estimators make parameter estimation using a different distribution or arbitrary function when the error distribution is different from Gaussian (skewed,



heavy long-tailed, excess kurtosis, etc.). In this respect, LS and M-Estimator methods can be said to be MLE estimators (Rousseeuw and Leroy 1987; Andersen 2008). In this study, Huber and Tukey M-Estimators, which are the most popular M-Estimators, will be discussed.

### 3.1.1 Huber M-Estimator

Huber (1964) proposed an M-Estimator that consists of Objective and Influence Function, which is the most popular robust estimation method. The most important characteristic of Huber Objective Function is that it acts like Gaussian distribution in center and Laplace distribution in tails (Hogg 1979). Objective Function, Influence Function and Weight Function of Huber M-Estimator are given in Table 1.

In Table 1,  $k$  is tuning constant and default value is 1.345 for 95% efficiency under normal distribution. According to the weight function, weights, the observations in the center of the distribution are equal and 1, while inversely proportional to the absolute value of the observations as they move away from the center (Fox and Weisberg 2002: 3). The graphs of Objective, Influence and Weight Function of Huber M-Estimator can be seen in Fig. 2a.

### 3.1.2 Tukey’s (Bisquare) M-Estimator

Tukey M-Estimator or Tukey’s Bi-Weight (Bisquare) based on weight function was first proposed by Beaton ve Tukey (1974). Objective Function, Influence Function and Weight Function of Tukey M-Estimator are given in Table 2.

**Table 1** Objective function, influence function and weight function of Huber M-Estimator

$\rho(r) = \begin{cases} \frac{1}{2}r^2 & ,  r  < k \\ k r  - \frac{1}{2}k^2 & ,  r  \geq k \end{cases}$	$\psi(r) = \begin{cases} r & ,  r  < k \\ k \text{sign}(r) & ,  r  \geq k \end{cases}$	$w(r) = \begin{cases} 1 & ,  r  < k \\ \frac{k}{ r } & ,  r  \geq k \end{cases}$
Objective function	Influence function	Weight function

**Table 2** Objective function, influence function and weight function of Tukey M-Estimator

$\rho(r) = \begin{cases} \frac{k^2}{6} \left\{ 1 - \left[ 1 - \left( \frac{r}{k} \right)^2 \right]^3 \right\} & ,  r  < k \\ \frac{k^2}{6} & ,  r  \geq k \end{cases}$	$\psi(r) = \begin{cases} r \left[ 1 - \left( \frac{r}{k} \right)^2 \right]^2 & ,  r  < k \\ 0 & ,  r  \geq k \end{cases}$	$w(r) = \begin{cases} \left[ 1 - \left( \frac{r}{k} \right)^2 \right]^2 & ,  r  < k \\ 0 & ,  r  \geq k \end{cases}$
Objective function	Influence function	Weight function

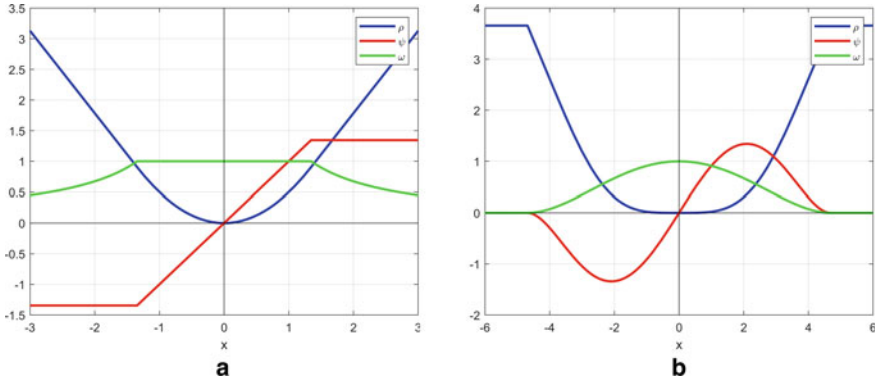


Fig. 2 a Huber M-estimator and b Tukey M-estimator

As the Huber M-Estimator, Tukey M-Estimator also has a tuning constant  $k$ , which its default value is 4.685 for 95% efficiency under normal distribution. The graphs of Objective, Influence and Weight Function of Tukey M-Estimator can be seen in Fig. 2b.

### 4 Pearson Distribution System

The Pearson Differential Equation (PDE) was first proposed by Karl Pearson (1895):

$$\frac{f'(x)}{f(x)} = \frac{d \ln f(x)}{dx} = \frac{x - a}{c_0 + c_1x + c_2x^2} \tag{11}$$

The solution of Eq. (11) defines Pearson Distribution Family (PDF), which consists of 13 different distributions with 3 main types and 10 transitional types. In Eq. (11), the parameter  $a$  is the mode of distribution and the parameters  $a, c_0, c_1, c_2, \dots$  can define type of the distribution uniquely. The function  $C(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots$ , in the dominator of the differential equation, defining as a polynomial allows the Method of Moments (MoM) can be used in estimating the unknown parameters of the differential equation (Şehirlioğlu and Dündar 2014).

If we solve Eq. (11), the consecutive moment equation is:

$$nc_0\mu'_{n-1} - \{(n + 1)c_1 - a\}\mu'_n - \{(n + 2)c_2 + 1\}\mu'_{n+1} = 0 \tag{12}$$

Substituting the values for  $n = 0,1,2,3$  in Eq. (12) for  $\mu'_1 = 0$ :

$$\begin{aligned}
 c_0 &= -\frac{\mu_2(4\mu_2\mu_4 - 3\mu_2^2)}{10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3} = -\frac{\sigma^2(4\beta_2 - 3\beta_1)}{10\beta_2 - 12\beta_1 - 18} \\
 c_1 = a &= -\frac{\mu_3(\mu_4 - 3\mu_3^2)}{10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3} = -\frac{\sigma\sqrt{\beta_1}(\beta_2 + 3)}{10\beta_2 - 12\beta_1 - 18} \\
 c_2 &= -\frac{2\mu_4\mu_2 - 3\mu_3^2 - 6\mu_2^3}{10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3} = -\frac{2\beta_2 - 3\beta_1 - 6}{10\beta_2 - 12\beta_1 - 18}
 \end{aligned}
 \tag{13}$$

where  $\beta_1 = \mu_3^2/\mu_2^3$  (Skewness) and  $\beta_2 = \mu_4/\mu_2^2$  (Kurtosis) parameters. The roots of  $C(x)$  determine the type of PDS. The three main types of PDS:

- Type I (PI): Roots are real and different signs.
- Tip IV (PIV): Roots are complex.
- Tip VI (PVI): Roots are real and same sign.

Another method that can be used to determine the distribution types is the Kappa ( $\kappa$ ) criterion. The coefficient of Kappa is a statistics obtained by using the discriminant of the  $C(x)$  function (Elderton 1906; Hald 2008; Fiori and Zenga 2009; Nagahara 2008). The discriminant of the  $C(x)$  and  $\kappa$  coefficient (Pearson 1901) can be written as follows:

$$\Delta = c_1^2 - 4c_0c_2 \tag{14}$$

$$\kappa = \frac{c_1^2}{4c_0c_2} = \frac{\beta_1(\beta_2 + 3)^2}{4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)} \tag{15}$$

The distributions according to  $\kappa$  criteria are given in Table 3.

An important case for PDS is the origin of the distribution must be mode point ( $a = 0$ ). For this reason, substitution  $X = x - a$  in Eq. (11), the PDE can be rewritten as follows:

$$\frac{df(X)}{dX} = \frac{Xf(X)}{C_0 + C_1X + C_2X^2} = \frac{(x - a)f(x)}{c_0 + c_1(x - a) + c_2(x - a)^2} \tag{16}$$

By solving Eq. (16), one can easily obtain the parameters which mode point coincides with the origin. The new parameters, which is  $a = 0$ , can be written in

**Table 3** Main types of PDS

$\Delta$	$\kappa$ Statistics	$c_0$	$c_1$	$c_2$	Roots	Type
$\Delta > 0$	$\kappa < 0$	$c_0 \neq 0$	$c_1 \neq 0$	$c_2 \neq 0$	Real	Type I
$\Delta > 0$	$\kappa > 1$	$c_0 \neq 0$	$c_1 \neq 0$	$c_2 \neq 0$	Real	Type VI
$\Delta < 0$	$0 < \kappa < 1$	$c_0 \neq 0$	$c_1 \neq 0$	$c_2 \neq 0$	Complex	Type IV

(Şehirlioğlu and Dündar 2014: 16)

terms of original parameters as follows:

$$\begin{aligned}
 c_2 &= C_2 \\
 2ac_2 + c_1 &= a(2c_2 + 1) = C_1 \\
 c_2a^2 + c_1a + c_0 &= c_0 + a^2(1 + c_2) = C_0
 \end{aligned}
 \tag{17}$$

### 4.1 Pearson Type I Distribution (PI)

Pearson Type I Distribution (PI) is one of the main types of PDF. The roots of  $C(x)$  must be different signs and the parameters should be  $c_2 > 0$  and  $c_0 < 0$ . The PI is also known as Beta Distribution. The pdf of PI:

$$f(x) = K(r_1 - x)^{m_1}(x + r_2)^{m_2}, \quad r_2 < x < r_1 \tag{18}$$

and

$$K = \frac{1}{B(m_1 + 1; m_2 + 1)(r_1 + r_2)^{m_1+m_2+1}} \tag{19}$$

where  $B(x, y)$  is a Beta Function,  $r_1$  is scale,  $r_2$  is location and  $m_1, m_2$  skewness and kurtosis parameters and  $K$  is the constant of normalization to make sure  $\int f(x)dx = 1$ . For the integral constant  $K$ , see more details of Pearson (1895), Nagahara (2008) and Elderton (1953). Figure 3 shows pdfs of PI with different skewness and kurtosis parameters.

### 4.2 Pearson Type IV Distribution (PIV)

Pearson Type IV Distribution (PIV) is the hardest distribution in PDF. For the existence of PIV, the roots of  $C(x)$  must be complex. The pdf of PIV:

$$f(x) = K[(x + r)^2 + s^2]^m e^{v \arctan \tau}, \quad -\infty < x < \infty \tag{20}$$

where  $K = \frac{s^{-2m-1}}{\exp(\frac{v\pi}{2}) \int_{-\pi/2}^{\pi/2} (\cos \theta)^{-2m-2} \exp(-v\theta) d\theta}$ ,  $m = 1/2c_2, v = (a+r)/sc_2, r = real(r_1)$

and  $s = imag(r_1)$ . Figure 4 shows pdfs of PIV with different skewness and kurtosis parameters.

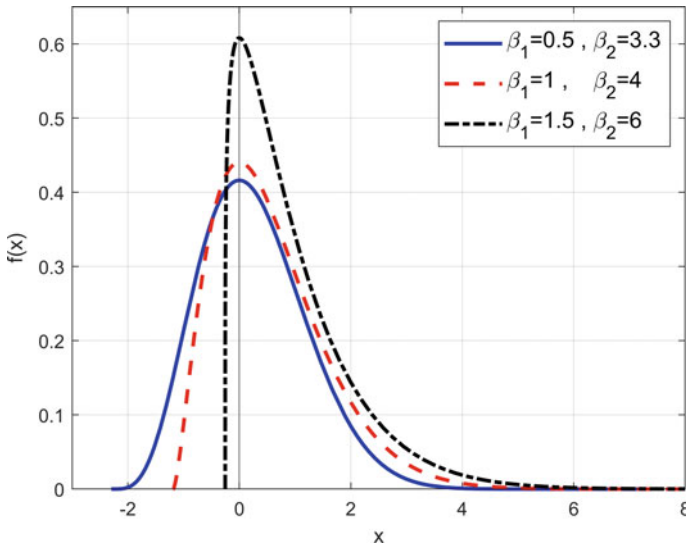


Fig. 3 Probability density functions of PI with different skewness and kurtosis

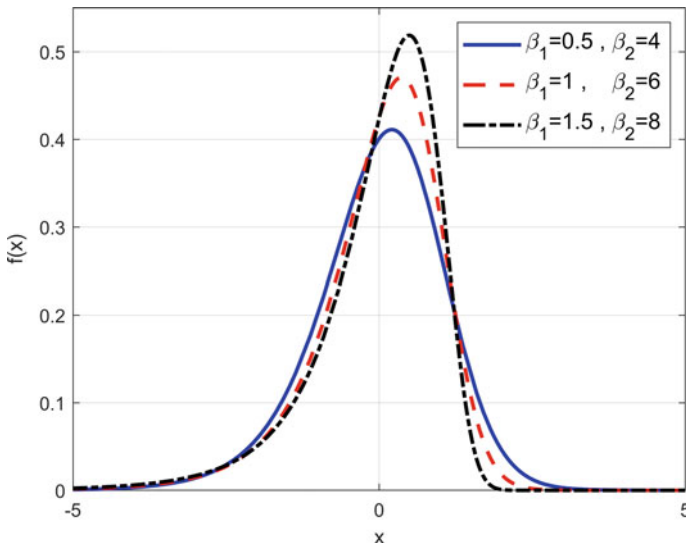
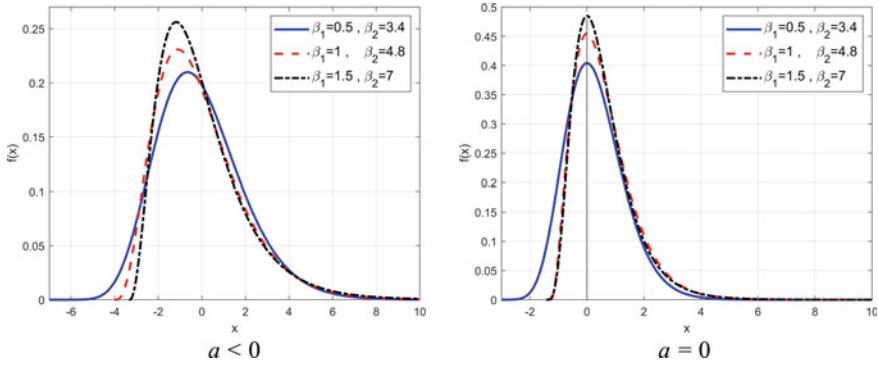


Fig. 4 Probability density functions of PIV with different skewness and kurtosis

### 4.3 Pearson Type VI Distribution (PVI)

Pearson Type VI Distribution (PVI) has the same sign of roots of  $C(x)$ . For  $-r_2 < -r_1 < 0$  right skewed PVI:



**Fig. 5** Probability density functions of PVI with different skewness and kurtosis

$$f(x) = K(x + r_1)^{m_1}(x + r_2)^{m_2}, \quad -r_1 < x < \infty \tag{21}$$

and

$$K = \frac{1}{B(-m_1 - m_2 - 1; m_1 + 1)(r_2 - r_1)^{m_1+m_2+1}}. \tag{22}$$

where  $r_1$  is scale,  $r_2$  is location and  $m_1, m_2$  skewness and kurtosis parameters. Use Eq. (16) for the PVI with mode at origin can be written as follows:

$$f(X) = K^*(X + R_1)^{M_1}(X + R_2)^{M_2}, \quad R_1 < R_2 < X < \infty \tag{23}$$

Figure 5 shows pdfs of PVI with different skewness and kurtosis parameters when  $a < 0$  and  $a = 0$ .

## 5 Pearson Differential Equation as a Influence Function

The similarity between the Influence Function (IF) and the PDE, different IFs can be defined for the dynamic parameters of the PDE. Thus, regression parameter estimates can be made by using the Weight Function. Dzhun' (2011) shows the similarity between IF and PDE as follows:

$$\psi(x) = \frac{d\rho(x)}{dx} = \frac{d[-\ln f(x)]}{dx} = \frac{f'(x)}{f(x)} = \frac{x - a}{c_0 + c_1x + c_2x^2} \tag{24}$$

Using Eq. (24), different IFs can be easily obtained. (Wiśniewski 2014). The definition of Weight Function is  $w(x) = \psi(x)/x$ :

$$w(x) = \frac{x - a}{x(c_0 + c_1x + c_2x^2)} \quad (25)$$

The value of the weight function is usually closely related to the mode point of the distribution. If the data have skewness and excess kurtosis, the Weight Function should take its maximum value at the mode point. In such cases, Eq. (16) can be used for coincides to mode point and the origin. IF, where the mode point at the origin:

$$\psi(X) = \frac{X}{C_0 + C_1X + C_2X^2} \quad (26)$$

and the Weight Function:

$$w(X) = \frac{1}{C_0 + C_1X + C_2X^2} \quad (27)$$

In this study, we only consider PVI for estimating regression parameters. For this purpose, Objective, IF and Weight Function will only be obtained for PVI. Consider Eq. (23), if the constant term removed, the Objective Function of PVI:

$$\begin{aligned} f(X) &\propto (X + R_1)^{M_1}(X + R_2)^{M_2} \\ \rho(X) &= -\ln f(X) = -M_1 \ln(X + R_1) - M_2 \ln(X + R_2) \end{aligned} \quad (28)$$

The Objective Function of PVI provides flexibility in terms of functions that will be minimized, due to its location, scale, skewness and kurtosis parameters compared with other robust methods. Based on Eq. (28), the IF and Weight Function of PVI:

$$\psi(X) = \frac{d\rho(X)}{dX} = \frac{d[-\ln f(X)]}{dX} = -\frac{M_1}{X + R_1} - \frac{M_2}{X + R_2} \quad (29)$$

and

$$w(X) = \frac{\psi(X)}{X} = \frac{M_1R_2 + M_2R_1 - (M_1 + M_2)X}{X(X + R_1)(X + R_2)} \quad (30)$$

Figure 6 shows Objective, Influence and Weight Functions of PVI with different skewness and kurtosis parameters.

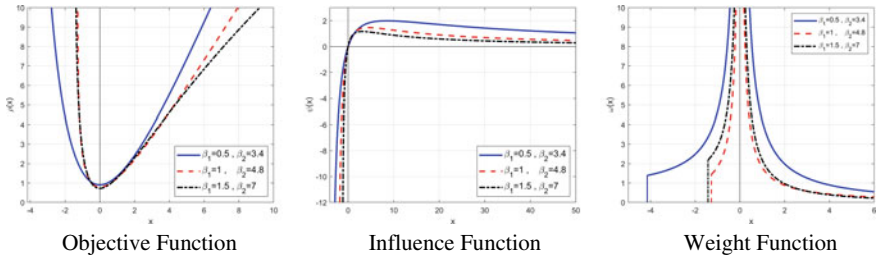


Fig. 6 Objective, Influence and Weight Functions of PVI

## 6 Real Data Examples and Simulation Study

### 6.1 Real Data Examples

In the Real Data example, two different data sets are analyzed for the estimation method based on the weight function of the proposed PVI function. The first data set is Education Expenditure, which is commonly used in the robust literature, and the second data set is Industrial Production Index, Unemployment Rate and CPI values in Turkey. For the data sets, goodness of fit of the PVI function is applied, and standard errors of the regression parameters are obtained. Box-plot and histogram of residuals for data sets are drawn for residuals obtained from LS.

#### 6.1.1 Education Expenditure Data

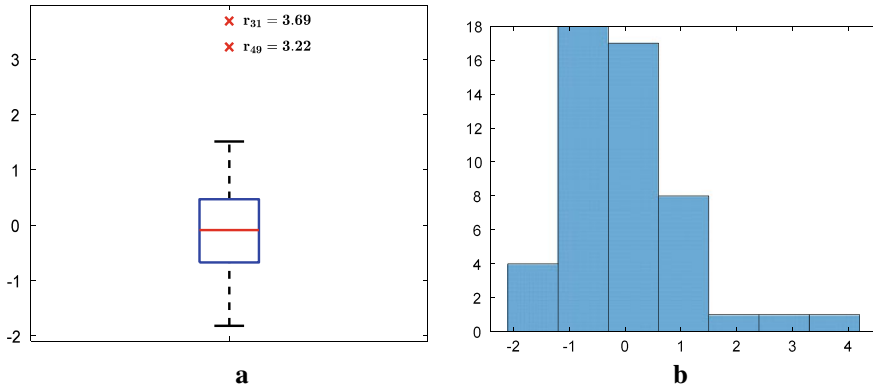
The data set published by Chatterjee and Price (1977) discussed Education Expenditures in 50 states in the United States. Information about the variables as follows: Average education expenditure per capita at public school in a state in 1975 (response variable)

- Number of residents residing in urban areas in 1970 ( $x1000$ ),
- Per capita personel income in 1973,
- Number of residents under 18 years of age in 1974 ( $x1000$ ).

In the data set, the value of 31st data of the response variable has been replaced with the value 400 instead of 212 due to a recording or transferring error. After that, PVI parameters were estimated and regression analysis was performed. When Figure 7a is examined, it can be seen that the 31st and 49th data points have high standardized residuals (3.69 and 3.22). Anderson-Darling goodness of fit test is applied to the residuals, and the test value is obtained as 1.046 ( $p = 0.000$ ), and it is determined that the distribution of the residuals does not come from normal distribution. Also, the skewness is 1.10, and kurtosis is 5.30 of residuals obtained from LS.

In Table 4, the PVI method gives similar results with other estimation methods even at high skewness and kurtosis values. Although Huber and Tukey M-Estimation





**Fig. 7** Box plot and histogram of education expenditure data

**Table 4** Result of education expenditure data

Estimation method	Results	Constant	Number of residents residing in urban areas (x1000)	Variables	
				Per capita personel income	Number of residents under 18 years of age (x1000)
LS	Estimate	-452.203	0.001	0.063	1.346
	Standard Error	146.891	0.061	0.014	0.375
Huber M-estimator	Estimate	-340.860	0.044	0.053	1.067
	Standard error	126.629	0.053	0.012	0.323
Tukey M-estimator	Estimate	-243.762	0.074	0.045	0.820
	Standard Error	127.490	0.053	0.012	0.326
PVI	Estimate	-307.984	0.009	0.061	0.909
	Standard error	72.326	0.030	0.007	0.185

methods give very low or 0 weight when residuals increase, the PVI method analyzed high standardized residuals as a part of the data set and weighed all residuals. Thus, when making regression diagnostic, when there is an observation that seems to be an outlier but is known to belong to the data set and even if the skewness and kurtosis values are high, the PVI method gives results at least as well as other estimation methods.

**6.1.2 Economical Data**

For the Economic data set, between January 2015 and February 2020, the values of 62 month Industrial Production Index (2015 = 100), Unemployment Rate and

Consumer Price Index (2003 = 100, response) are considered. In this example, by considering the variables commonly used in economic analysis in Turkey, we have been focused on the importance of determining the exact distribution of the data. When Fig. 8a is examined, it can be seen that observations 45th, 46th, 47th, 48th, 49th, 52th and 57th have high standardized residuals. In addition, if the histogram of the residuals (Fig. 8a) is examined, it is seen that the data are right skewed ( $\beta_1 = 1.08$ ) and have excess kurtosis ( $\beta_2 = 5.04$ ). As a result of the Anderson–Darling test, the test statistic is 1.782 ( $p = 0.005$ ) and the distribution of the residuals is not normal.

According to the regression analysis results in Table 5, PVI method gives very similar results to other methods. Although the skewness and kurtosis values of the data are high and there are high standardized residuals, the coefficients were obtained significantly.

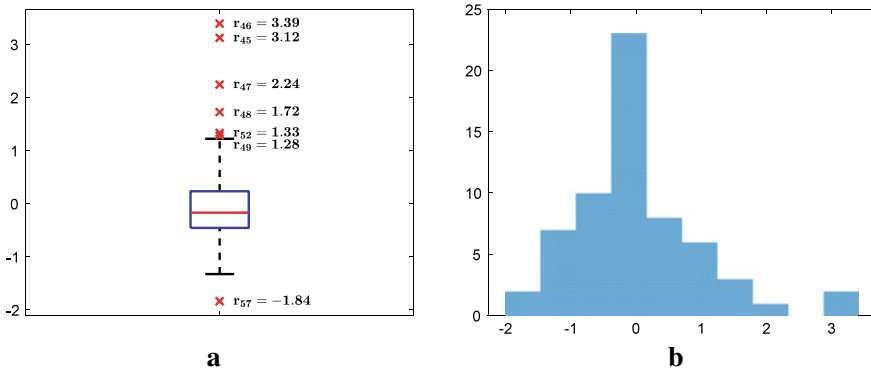


Fig. 8 Box plot and histogram of economical data

Table 5 Result of economical data

Estimation method	Results	Variables		
		Constant	Unemployment rate	Industrial production index (2015 = 100)
LS	Estimate	-29.615	0.256	1.143
	Standard error	8.052	0.077	0.338
Huber M-estimator	Estimate	-28.035	0.250	1.012
	Standard error	6.716	0.064	0.282
Tukey M-estimator	Estimate	-17.680	0.236	0.644
	Standard error	4.722	0.045	0.128
PVI	Estimate	-25.315	0.247	0.781
	Standard error	8.383	0.080	0.352

## 6.2 Simulation Study

In the simulation study, LS, Huber M-Estimator, Tukey M-Estimator and proposed method based on PVI are compared with different scenarios. Sample sizes 30, 50, 100 and 500 are used and  $M=1000$  replications are simulated. Data are generated multivariate linear regression using following model;

$$y = 2 + 2X_1 + 2X_2 + 2X_3 + 2X_4 + 2X_5 + \varepsilon, \quad X_i \sim N(0, 1) \quad (31)$$

Error term is generated 13 different scenarios and 7 distributions with 2 symmetrical and 5 asymmetrical distributions. Predetermined scenarios of the error term are as follows:

**Scenario 1.**  $\varepsilon \sim N(0, 1)$ , Standard Normal Distribution (PXIII).

**Scenario 2.**  $\varepsilon \sim t(1)$ , (PVII).

**Scenario 3.**  $\varepsilon \sim t(10)$ , (PVII).

**Scenario 4.**  $\varepsilon \sim Exp(1)$ , (PX).

**Scenario 5.**  $\varepsilon \sim Exp(10)$ , (PX).

**Scenario 6.**  $\varepsilon \sim Gamma(1, 5)$ , (PIII).

**Scenario 7.**  $\varepsilon \sim Gamma(2, 5)$ , (PIII).

**Scenario 8.**  $\varepsilon \sim \chi^2(1)$ , (PIII).

**Scenario 9.**  $\varepsilon \sim \chi^2(5)$ , (PIII).

**Scenario 10.**  $\varepsilon \sim F(2, 10)$ , (PVI).

**Scenario 11.**  $\varepsilon \sim F(10, 10)$ , (PVI).

**Scenario 12.**  $\varepsilon \sim Weibull(1, 1)$ ,

**Scenario 13.**  $\varepsilon \sim Weibull(2, 2)$ .

In each scenario, residuals are obtained from LS estimation method and we choose response and explanatory variables, which is suitable for PVI. Initial weights are calculated for Huber M-Estimator and Tukey M-Estimator using the residuals obtained from LS and for the PVI method mode point must coincide at origin. We use Iteratively Re-Weighted Least Squares (IRWLS) for estimating regression parameters and calculate Total Absolute Bias (TAB) and Mean Squared Error (MSE) for comparing estimation methods. The calculating steps of IRWLS as follows (Fox and Weisberg 2002; Maronna et al. 2006; Hogg 1979):

1. Choose a suitable estimation method (usually LS) and estimate regression parameters.

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

2. Calculate residuals using the estimated regression parameter.

$$\mathbf{e} = \mathbf{y} - \mathbf{X} \hat{\theta}.$$

3. Estimate robust standard deviation (Usually MAD).

4. Calculate studentized residuals using MAD and use tuning constant if exist.

$$r_i = e_i / \left( ks \sqrt{(1 - h_i)} \right)$$

5. Choose a weight function from Table 1c, Table 2c or Eq. (30).
6. Estimate new regression parameters using Weighted LS method.

$$\hat{\theta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

7. Repeat 2–6 until converge:

$$\left\| \hat{\theta}^i - \hat{\theta}^{i-1} \right\| / \left\| \hat{\theta}^i \right\| < 10^{-6} \text{ or } i > 30.$$

While comparing the estimation methods, we can calculate TAB and MSE using regression parameters as follows (Wiśniewski 2014);

$$TAB = \sum_{j=0}^5 \left| \frac{1}{M} \sum_{i=1}^M \hat{\theta}_{ij} - \theta_{ij} \right|$$

$$MSE = \frac{1}{M} \sum_{i=1}^M (\hat{\theta}_i - \theta)^T (\hat{\theta}_i - \theta).$$

Table 6 shows TAB and MSE values for each estimation method and sample sizes when the error term follows both symmetric and asymmetrical distribution. Considering these values, LS method gives the best results when the distribution of the error term is normal distribution. However, when the distribution of the error term is *Student t* distribution, which has heavy and long-tailed than the normal distribution, Huber and Tukey M-Estimators give better results than both LS and PVI. For some distributions in simulations, PVI is not suitable in large sample sizes. In every case of the error term follows asymmetrical distribution, LS, Huber M and Tukey M-Estimators give worse results than PVI since they do not consider skewness and excess kurtosis of data set. The PVI weight function ensures a flexible structure according to the parameters of the distribution.

According to the simulation results, in the presence of skewness and excess kurtosis in the data, the PVI method has the best TAB and MSE values since the weight function takes these parameters into account. However, when the distribution of errors is symmetrical, the PVI method gives worse than other methods.

## 7 Conclusion

Researchers use many estimation methods in Robust Regression Analysis. The most important and most widely used robust estimation method is M-Estimators. However, traditional M-Estimators do not consider the anomalies (heavy-long tail, skewness

**Table 6** Simulation results

Simulation study		Symmetrical distributions										Asymmetrical distributions									
Sample size	Method	Criteria	Normal (0,1)	$t_1$	$t_{10}$	Exp (1)	Exp (10)	Gamma (1,5)	Gamma (2,5)	$\chi^2_1$	$\chi^2_5$	F (2,10)	F (10,10)	Weibull (1,1)	Weibull (2,2)						
30	LS	TAB	<b>0.0303</b>	15.621	0.0922	10.178	101.781	50.819	101.022	10.416	50.487	13.022	12.908	10.187	17.646						
		MSE	<b>0.2589</b>	230.663	0.3453	12.658	12.657	315.408	1138	16.044	274.566	21.606	18.987	12.668	32.608						
	Huber M-Estimator	TAB	0.0808	0.2966	<b>0.0617</b>	0.8284	82.837	41.904	89.912	0.7200	46.236	0.9670	11.268	0.8270	16.791						
		MSE	0.2902	18.084	<b>0.3267</b>	0.8349	83.4.860	210.001	90.4.979	0.7256	227.205	11.523	13.736	0.8381	29.843						
	Tukey M-Estimator	TAB	0.1278	<b>0.1276</b>	0.1117	0.7569	75.692	38.179	85.701	0.6027	44.150	0.8499	10.614	0.7502	16.346						
		MSE	0.3437	<b>11.645</b>	0.3769	0.7206	720.546	180.047	831.000	0.5310	209.838	0.9246	12.138	0.7172	28.650						
	PVI	TAB	0.1651	0.0722	0.1427	<b>0.6891</b>	<b>68.899</b>	<b>35.089</b>	<b>81.770</b>	<b>0.5276</b>	<b>43.171</b>	<b>0.7715</b>	<b>10.215</b>	<b>0.6917</b>	<b>16.053</b>						
		MSE	0.3957	13.141	0.4490	<b>0.6547</b>	<b>654.653</b>	<b>168.198</b>	<b>789.426</b>	<b>0.4956</b>	<b>203.966</b>	<b>0.8434</b>	<b>11.770</b>	<b>0.6737</b>	<b>28.109</b>						
	50	LS	TAB	<b>0.0171</b>	0.8723	0.0717	10.124	101.240	50.742	100.975	10.230	50.890	12.630	12.662	10.232	17.723					
			MSE	<b>0.1366</b>	36.184	0.1842	11.453	1145.326	283.978	1076	12.796	265.389	18.317	16.900	11.371	31.984					
Huber M-Estimator		TAB	0.0644	0.2315	<b>0.0330</b>	0.8290	82.895	41.691	91.359	0.6916	46.795	0.9635	11.113	0.8441	16.951						
		MSE	0.1539	0.7217	<b>0.1749</b>	0.7662	766.194	191.035	875.446	0.5761	223.527	10.395	12.744	0.7636	29.370						
Tukey M-Estimator		TAB	0.0933	<b>0.0697</b>	0.0682	0.7588	75.879	38.373	88.001	0.5603	45.095	0.8466	10.521	0.7765	16.646						
		MSE	0.1760	<b>0.5445</b>	0.1928	0.6577	657.656	164.799	813.066	0.4031	209.263	0.8245	11.409	0.6565	28.379						
PVI		TAB	0.1438	0.0867	0.1268	<b>0.6682</b>	66.823	<b>33.031</b>	<b>80.767</b>	<b>0.4618</b>	<b>42.211</b>	<b>0.7286</b>	<b>0.9710</b>	<b>0.6834</b>	<b>15.990</b>						
		MSE	0.2186	0.5659	0.2452	<b>0.5368</b>	<b>536.811</b>	<b>132.652</b>	<b>712.947</b>	<b>0.3061</b>	<b>187.777</b>	<b>0.6531</b>	<b>10.071</b>	<b>0.5333</b>	<b>26.575</b>						
100		LS	TAB	<b>0.0129</b>	-	0.0525	10.139	<b>101.394</b>	50.372	100.787	10.223	50.460	12.355	12.545	10.082	17.708					
			MSE	<b>0.0648</b>	-	0.0802	10.663	1066.307	261.131	1032	11.186	255.428	16.233	15.769	10.613	31.551					
	Huber M-Estimator	TAB	0.0445	-	<b>0.0228</b>	0.8412	84.116	41.620	91.534	0.6817	46.692	0.9512	11.150	0.8385	17.018						
		MSE	0.0711	-	<b>0.0755</b>	0.7300	729.949	179.442	853.479	0.4995	218.271	0.9471	12.339	0.7282	29.212						

(continued)

**Table 6** (continued)

Simulation study		Symmetrical distributions					Asymmetrical distributions							
<i>Tukey</i>	<i>TAB</i>	0.0584	–	0.0417	0.7853	78.534	38.809	88.876	0.5540	45.643	0.8509	10.673	0.7850	16.829
	<i>MSE</i>	0.0764	–	0.0789	0.6420	641.947	157.784	808.076	0.3442	208.684	0.7636	11.334	0.6413	28.590
	<i>PVI</i>	0.1177	–	0.1180	<b>0.6485</b>	<b>64.850</b>	<b>32.109</b>	<b>79.096</b>	<b>0.4249</b>	<b>41.509</b>	<b>0.6935</b>	<b>0.9655</b>	<b>0.6432</b>	<b>15.901</b>
<i>LS</i>	<i>MSE</i>	0.1058	–	0.1130	<b>0.4534</b>	<b>453.435</b>	<b>112.153</b>	<b>659.232</b>	<b>0.2209</b>	<b>174.865</b>	<b>0.5346</b>	<b>0.9347</b>	<b>0.4505</b>	<b>25.913</b>
	<i>TAB</i>	–	–	0.0334	10.067	100.666	50.170	100.442	10.044	50.192	12.143	12.338	10.045	17.726
	<i>MSE</i>	–	–	0.0153	10.105	1010.528	252.585	1005	10.205	250.864	14.840	15.199	10.113	31.369
<i>Huber</i>	<i>TAB</i>	–	–	<b>0.0189</b>	0.8418	84.177	42.093	91.909	0.6648	46.733	0.9429	11.027	0.8438	17.115
	<i>MSE</i>	–	–	<b>0.0144</b>	0.7115	71.460	177.435	841.378	0.4464	217.276	0.8946	12.119	0.7113	29.273
	<i>TAB</i>	–	–	0.0254	0.7961	79.612	39.860	90.334	0.5433	46.163	0.8541	10.674	0.7995	17.061
<i>M-Estimator</i>	<i>MSE</i>	–	–	0.0146	0.6382	638.232	159.041	813.301	0.2987	211.924	0.7340	11.353	0.6381	29.098
	<i>TAB</i>	–	–	0.0934	<b>0.6209</b>	<b>62.092</b>	<b>31.087</b>	<b>78.784</b>	<b>0.3742</b>	<b>41.184</b>	<b>0.6426</b>	<b>0.9308</b>	<b>0.6250</b>	<b>15.993</b>
	<i>MSE</i>	–	–	0.0266	<b>0.3910</b>	<b>390.964</b>	<b>97.447</b>	<b>619.100</b>	<b>0.1474</b>	<b>169.547</b>	<b>0.4249</b>	<b>0.8639</b>	<b>0.3925</b>	<b>25.656</b>

and kurtosis) found in the data set. In this study, based on the similarity between PDE and IF, when the distribution of the data follows the PVI distribution, the performance of the M-Estimator based on the weighting function of the PVI distribution is considered. If the error term has symmetrical distributions, the proposed method gives worse performance than other methods. However, in cases where the distribution of the error term is asymmetric ( $\beta_1 > 0$ ) and excess kurtosis ( $\beta_2 > 3$ ), the weight function of the PVI distribution has a better performance than other M-Estimators since it contains these parameters. Also, due to the heavy and long tails (see Appendix) of the PVI, it performs well in asymmetrical distributions with light tails. Asymmetrical distributions used in simulations have a lighter tail than PVI.

The performance of the PVI has been analyzed by performing a simulation study on 13 scenarios with 7 different distributions with two different real-world data.

## Appendix

### *Tail Properties of a Distribution*

The Tail Function of a distribution  $G$  can be defined as follows:

$$\overline{G}(x) = G(x, \infty), \quad x \in R$$

Considering the tail function or pdf, it can be determined whether a distribution has a Heavy-Tailed, Fat-Tailed or Long-Tailed by considering the following three conditions (Bryson 1974; Foss et al. 2011).

#### **Condition 1. Heavy-Tailed Distributions**

$G$  can be defined as Heavy-Tailed distribution, it must be satisfied;

$$\int_R e^{\lambda x} G(dx) = \infty, \quad \forall \lambda > 0$$

If the Heavy Tailness has written for pdf;

$$\lim_{x \rightarrow \infty} \sup g(x)e^{\lambda x} = \infty, \quad \forall \lambda > 0$$

#### **Condition 2. Long-Tailed Distributions**

$G$  can be defined as Long-Tailed distribution, it must satisfy;

$$\lim_{x \rightarrow \infty} \frac{g(x + \lambda)}{g(x)} = 1, \quad \forall \lambda > 0$$

A Long-Tailed distribution is subclass of Heavy-Tailed distributions. By using Tail Function;

$$\lim_{x \rightarrow \infty} \overline{G}(x + \lambda) \sim \overline{G}(x), \quad \forall \lambda > 0$$

**Condition 3. Fat-Tailed Distributions**

$G$  can be defined as Fat-Tailed distribution, it must be satisfy;

$$\lim_{x \rightarrow \infty} P(X > x) \sim x^{-\lambda}, \quad \forall \lambda > 0$$

where  $P(X > x) = \overline{G}(x)$ .

In this study, we investigated tail properties of PVI. Consider Eq. (23);

**Proof of Condition 1**

By using pdf of PVI and Condition 1, one can easily calculate;

$$\lim_{x \rightarrow \infty} e^{\lambda x} f(x) = \infty$$

Thus PVI is a Heavy Tailed Distribution

**Proof of Condition 2**

For any  $\lambda > 0$ , we obtain the equation  $f(x + \lambda) = (x + r_1 + \lambda)^{m_1} (x + r_2 + \lambda)^{m_2}$ .

By using Condition 2;

$$\lim_{x \rightarrow \infty} \frac{f(x + \lambda)}{f(x)} = 1$$

Thus, PVI is a Long-Tailed Distribution.

**Proof of Condition 3**

The Tail Function of PVI;



$$P(X > x) = 1 - \frac{x^{m_1+1} {}_2F_1 \left[ \begin{matrix} m_1 + 1 & -m_2 \\ m_1 + 2 \end{matrix}; -x \right]}{(m_1 + 1)B(-m_1 - m_2 - 1; m_2 + 1)}$$

where  ${}_2F_1 \left[ \begin{matrix} a & b \\ c \end{matrix}; z \right]$  is Gauss Hypergeometric Function. By using the Condition 3;

$$\lim_{x \rightarrow \infty} \bar{F}(x) = -\infty$$

Thus PVI is not a Fat-Tailed Distribution.

## References

- Agee WS, Turner RH (1978). Application of robust statistical methods to data reduction (No. ACD-TR-65). National range operations directorate white sands missile range N Mex analysis and computation div
- Allende H, Frery AC, Galbiati J, Pizarro L (2006) M-estimators with asymmetric influence functions: the distribution case. *J Stat Comput Simul* 76(11):941–956
- Andersen R (2008). Modern methods for robust regression (No. 152). Sage, Thousand Oaks
- Andrews DF (1974) A robust method for multiple linear regression. *Technometrics* 16(4):523–531
- Balkema G, Embrechts P (2018) Linear Regression for Heavy Tails. *Risks* 6(3):93
- Barnett V, Lewis T (1984) Outliers in statistical data (OSD)
- Beaton AE, Tukey JW (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* 16(2):147–185
- Bell RM (1980). An adaptive choice of the scale parameter for M-estimators (No. TR-3). Stanford Univ Ca Dept of Statistics, Stanford
- Box GE (1953) Non-normality and tests on variances. *Biometrika* 40(3/4):318–335
- Brascamp JW, Berg AV, Ee R (2004) Shared neural circuitry for switching between perceptual states and ocular motor states? *J vis* 4(8):255–255
- Bryson MC (1974) Heavy-tailed distributions: properties and tests. *Technometrics* 16(1):61–68
- Cantoni E, Ronchetti E (2006) A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *J Health Econ* 25(2):198–213
- Chatterjee S, Price B (1977) Regression analysis by example. Wiley, New York
- Chen C (2013) Distributed iteratively reweighted least squares and applications. *Stat Interface* 6(4):585–593
- Croux C, Reusseeuw PJ (1992) Time-efficient algorithms for two highly robust estimators of scale. In: Computational statistics. Physica, Heidelberg, pp 411–428
- Draper NR, Smith, H (2014). Applied regression analysis, vol 326. Wiley, Hoboken
- Dzhun' IV (2011) Method for diagnostics of mathematical models in theoretical astronomy and astrometry. *Kinemat Phys Celest Bodies* 27:260–264
- Elderton WP (1906) Frequency curves and correlation. Cambridge University, New York
- Elderton WP (1953) Frequency curves and correlation. Cambridge University, New York
- Fiori AM, Zenga M (2009) Karl Pearson and the origin of kurtosis. *Int Stat Rev* 77(1):40–50
- Foss S, Korshunov D, Zachary S (2011). An introduction to heavy-tailed and subexponential distributions, vol 6, pp 0090–6778. Springer, New York
- Fox J, Weisberg S (2002) Robust regression. *An R S Plus Companion Appl Regres* 91

- Hald A (2008) A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935. Springer Science & Business Media, Berlin
- Hampel FR (1974) The influence curve and its role in robust estimation. *J Am Stat Assoc* 69(346):383–393
- Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence functions. Wiley, New York
- Hampel FR (1968) Contribution to the theory of robust estimation. PhD thesis, University of California, Berkeley
- Harvey AC (1977) A comparison of preliminary estimators for robust regression. *J Am Stat Assoc* 72(360a):910–913
- Hinich MJ, Talwar PP (1975) A simple method for robust regression. *J Am Stat Assoc* 70(349):113–119
- Hogg RV (1979) Statistical robustness: one view of its use in applications today. *Am Stat* 33(3):108–115
- Huber PJ (1964) Robust version of a location parameter. *Ann Math Stat* 36:1753–1758
- Huber PJ, Ronchetti EM (1981) Robust statistics, ser. Wiley Ser Probab Math Stat New York, NY, USA, Wiley-IEEE 52:54
- Judge GG, Hill RC, Griffiths WE, Lütkepohl H, Lee TC (1988) Introduction to the theory and practice of econometrics (No. 330.015195 I61 1988). Wiley, Hoboken
- Khalil U, Ali A, Khan DM, Khan SA, Qadir F (2016) Efficient Uk'S re-descending M-estimator for robust regression. *Pak J Stat* 32(2)
- Kilany NM (2016) Weighted Lomax Distribution. *Springerplus* 5(1):1862
- Krasker WS, Kuh E, Welsch RE (1983) Estimation for dirty data and flawed models. *Handb Econ* 1:651–698
- Lenth RV (1977) Robust splines. *Commun Stat Theory Methods* 6(9):847–854
- Maronna RA, Martin D, Yohai RS (2006) Wiley series in probability and statistics. *Robust Stat Theory Methods* 404–414
- Mielke PW Jr, Johnson ES (1974) Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology. *Water Resour Res* 10(2):223–226
- Moghaddam MD, Liu J, Serota RA (2019) Implied and realized volatility: a study of distributions and the distribution of difference. *arXiv preprint [arXiv:1906.02306](https://arxiv.org/abs/1906.02306)*
- Mohebbi M, Nourijelyani K, Zeraati H (2007) A simulation study on robust alternatives of least squares regression. *J Appl Sci* 7(22):3469–3476
- Nagahara Y (2008) A method of calculating the downside risk by multivariate nonnormal distributions. *Asia Pac Financ Mark* 15(3–4):175–184
- Pearson K (1895) X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Philos Trans R Soc Lond A* 186:343–414
- Pearson, K. (1901). XI. Mathematical contributions to the theory of evolution.—X. Supplement to a memoir on skew variation. *Philos Trans R Soc Lond A (Containing Papers of a Mathematical or Physical Character)* 197(287–299):443–459
- Rasheed BA, Adnan R, Saffari SE, dano Pati K (2014) Robust weighted least squares estimation of regression parameter in the presence of outliers and heteroscedastic errors. *J Teknol* 71(1)
- Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection, vol 1. Wiley, New York
- Ryan TP (2008). *Modern regression methods*, vol 655. Wiley, New York
- Salazar JSI (2000) Detection schemes for synthetic-aperture radar imagery based on a beta prime statistical model.
- Şehirlioğlu AK, Dündar S (2014) Pearson Dağılışı Ailesi. İzmir: Ege Üniversitesi Basımevi
- Siegel AF (1982) Robust regression using repeated medians. *Biometrika* 69(1):242–244
- Staudte RG, Sheather SJ (2011) Robust estimation and testing, vol 918. Wiley, New York
- Stuart C (2011) Robust regression. Department of Mathematical Sciences, Durham University, Durham, p 169

- Sumarni C, Sadik K, Notodiputro KA, Sartono B (2017, March). Robustness of location estimators under t-distributions: a literature review. In: IOP conference series: earth and environmental science, vol 58, no 1. IOP Publishing, Bristol, p 012015
- Susanti Y, Pratiwi H (2014) M-estimation, S estimation, and MM estimation in robust regression. *Int J Pure Appl Math* 91(3):349–360
- Tukey JW (1962) The future of data analysis. *Ann Math Stat* 33(1):1–67
- Tukey JW (1960) A survey of sampling from contaminated distributions. *Contrib Probab Stat* 448–485
- TulupyeV A, Suvorova A, Sousa J, Zelterman D (2013) Beta prime regression with application to risky behavior frequency screening. *Stat Med* 32(23):4044–4056
- Wiśniewski Z (2014) M-estimation with probabilistic models of geodetic observations. *J Geodesy* 88(10):941–957
- Wood FS, Gorman JW (1971) Fitting equations to data: computer analysis of multifactor data for scientists and engineers. Wiley-Interscience, New York
- Wu LL (1985) Robust M-estimation of location and regression. *Sociol Methodol* 15:316–388
- Ye Y, Oluyede BO, Pararai M (2012) Weighted generalized beta distribution of the second kind and related distributions. *J Stat Econ Methods* 1(1):13–31
- Yulita T, Notodiputro KA, Sadik K (2018) M-estimation use bisquare, hampel, huber, and welsch weight functions in robust regression
- Zeigler RK, Ferris Y (1973) Half-life of Plutonium-241. *J Inorg Nucl Chem* 35(10):3417–3418

# Discrete Volatilities of Listed Real Estate Funds



Annah Gosebo, Donald Makhalemele, and Zinhle Simelane

**Abstract** The purpose of this article is to examine hedging strategies of South African real estate investment trusts using discrete volatility models. Prior studies have illustrated volatility hedging in bonds, commodities and equities appropriately illustrated by discrete volatility models, but not much has been done on real estate investment trusts, especially South African ones. This article uses both Autoregressive Conditional Heteroskedasticity and Generalized Autoregressive Conditional Heteroskedasticity family models to price discrete volatilities. The results show that information asymmetry, heterogeneity and lagging effects are inherent real estate investment trusts; therefore, volatility modelling should be done in a cautious manner. Thus, incorporating these factors in real estate investment trust hedging strategies should have a remarkable significance both in academia and in practice. The same findings apply equally both on in- and out-of-sample data.

**Keyword** Volatility · REITs · Hedging

## 1 Introduction

The analysis of volatility is relevant to various stakeholders in the capital market, including but not limited to investors, regulators, stock brokers and relevant firms. It can be inferred from Hoesli and Reka (2013) that volatility is a measure related with the risk and uncertainty connected with unexpected changes of an instrument price. The concept of volatility in stock pricing and/or returns therefore has a significant

---

A. Gosebo (✉) · D. Makhalemele · Z. Simelane  
School of Construction Economics & Management, WITS University, Braamfontein,  
Johannesburg 2050, South Africa  
e-mail: [727668@students.wits.ac.za](mailto:727668@students.wits.ac.za)

D. Makhalemele  
e-mail: [917211@students.wits.ac.za](mailto:917211@students.wits.ac.za)

Z. Simelane  
e-mail: [719648@students.wits.ac.za](mailto:719648@students.wits.ac.za)

impact on the hedging and risk management strategies that the investor might base its investment decisions on. Our research topic focuses on volatility analysis in real estate investment trusts (REITs) because of the nature and attributes that the REITs have when compared to the nature and behaviour of stocks which have been the topic of conversation that has been the most lacking in literature with respect to volatility hedging. Only a handful of recent literature has put the focus on REIT volatility that explores the complexities which result from underpinning of real and investment property. Our study will focus more so on listed property in the South African context, where the REIT legislation came to pass in 2013.

Due to the uniqueness of REITs to other elements in the capital market, such as common stocks, the topic of risk management has been a topic to be well explored in literature. However, due to the various models that exist to measure volatility—which represents the riskiness of a stock—it is difficult to choose one such perfect model to describe the volatility of returns while taking into consideration differences in market conditions as well as macroeconomic factors. Several studies over the years have shifted the focus to mechanisms that monitor and quantify market diversification in response to the increase in the growth of the globalization and diversification of the nature and behaviour of the economy (Ponta and Carbone 2018). The literature we have explored points out to deficiencies when it comes to the methodology as well as some overlooked variables that apply to volatility analysis, especially the one that involves REITs. Therefore, our problem will be based on exploring the various factors that continue to make risk management in the REIT industry.

The topic of REIT Volatility continues to be an interesting area of exploration in recent literature given the unique nature of REITs to the other investment vehicles that exist in the capital market. This is, according to Chaudhry et al. (2004), due to the organizational structure inherent in REITs which varies significantly from that of other common stocks, the difference between the REIT versus the common stock is also due to the close relationship inherent between the REIT and its underlying real estate property. With the stated differences in mind, the dynamics of volatility in REITs and in other common stocks are different and should, therefore, be investigated differently. Therefore, to effectively manage risk in REIT portfolios, it would be worthwhile to explore all factors that may affect the movement of returns.

While systematic risk has been widely researched by investigating industry-wide factors as well as fundamental economic factors, there is a gap in literature that supports further research into idiosyncratic risks inherent in the REITs industry (Cheng and Roulac 2007). Recent literature, such as Barkham and Geltner (1995), hypothesizes that if firm-specific factors of the underlying real estate asset influence average REIT returns, then the usage of the capital asset pricing model, their derivatives will leave a substantial amount of REIT volatility unexplained.

Then, the big question is how best can one model explain REIT volatility in the context of risk management? The contribution of this article is exactly answering that main question. To answer that main question, this article adopts the Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family models. The results show that the presence of information asymmetry, heterogeneity and the lagging effect is reflected

on the data sample, corresponding with the literature findings. Therefore, incorporating these factors in REIT hedging strategies should have a remarkable significance both in academia and in practice.

The balance of the article is as follows: Sect. 2 is on literature review, Sect. 3 is on methodology, Sect. 4 is data, Sect. 5 is on analysis and the last section concludes the article.

## 2 Literature Review

The subject of information asymmetry in REITs has gained more traction along the years in terms of literature as well as empirical tests. However, information asymmetry is said to be difficult to observe directly in the financial market; thus, empiricists have had to develop theoretical proxy variables to study it (Abdul-Baki 2013). Abdul-Baki (2013) further notes that there are not sufficient empirical studies that analyse the validity of the said proxies used to measure information asymmetry. His paper attempts to examine the validity of two proxies of information asymmetry that he found to be the most popular in literature. The first proxy that he used is the probability of informed trading and the second is the bid-ask spread (BAS). Among other things that contribute to information asymmetry includes (i) loan contract terms as a proxy for information asymmetry and (ii) bid-ask spreads as a proxy for Information asymmetry.

Deng et al. (2017) analysed the different outcomes of the presence in information asymmetry between externally managed REITs and internally managed REITs. The external REITs are characterized by their reliance on outside advisory and property management. Whereas, internal REITs have an in-house team managing the property and the company. They argue that although external REITs have exhibited poor performance due to the conflict-of-interest issue that arise between the outside advisors and the REIT shareholders; external REITs are more informationally efficient. In their analysis, they use loan contract terms as a proxy for information asymmetry, advancing that loan contract terms—such as interest rates, the number of covenants, loan maturity and whether there was collateral offered for the loan—reflect information asymmetry. Their claims are supported by earlier studies such as (Sufi 2007) who proposed that loan structure negotiated between the lender and borrower is influenced by the informational transparency of the borrower.

Extensive literature employs the Bid-Ask Spread (BAS) as a proxy for information asymmetry; this is due to the strong linkages it has demonstrated both empirically and theoretically with the level of information asymmetry in financial assets (Chung et al. 2017). Earlier research (Muller and Riedl 2002) centres its examination of information asymmetry risk in REITs on the bid-ask spread specifically due to the known fact that BAS is observable whereas other proxies of information asymmetry have been found to be difficult to observe empirically. Further, theoretical literature found strong associations between information asymmetry, BAS, and cost of capital in REITs. The bid-ask spread is defined empirically as the difference between the

price in which traders are willing to buy (bid) and sell (ask) a firm's securities. Furthermore, the bid-ask spread is an aggregate of three components of trading costs, namely, order processing, inventory holding as well as adverse selection costs. The latter component, adverse selection, has been the most popular in extent literature on information asymmetry due to the observation that adverse selection can yield high transaction costs as witnessed on the BAS due to information asymmetry (Muller and Riedl 2002).

The presence of heterogeneity is one of the reasons why hedging strategies in the REIT industry has difficulties. In essence, the concept of heterogeneity in the context of REITs refers to the diversification of the underlying property portfolio by property type and sector as well as geographically. Seiler et al. (1999) mentioned how REIT diversity differs from stock diversity where the latter varies only in terms of its market capitalization and industry, whereas the former varies in terms of more intrinsic property characteristics. These include the size of the property, the geographical location of the property and the asset type to name quite a few. Therefore, for a portfolio manager to enjoy the true benefits of diversification they should consider all the different ways in which the properties within the portfolio may vary to minimize risk and maximize returns.

Likewise, REIT managers could choose to diversify their REIT portfolio across the different asset classes such as residential, office, industrial as well as retail according to their performances. They could also diversify their portfolio by investing in properties across different geographical regions that could be in the form of countries or provinces within a country to minimize risk. Cheok et al. (2011) found that REITs diversified by their property types had no significant effect on risk; contrariwise, they found quite a significant impact on risk because of a geographically diversified property portfolio. Their main contribution to literature was to find the solution as to whether REITs with a homogenous portfolio yield better returns than diversified REIT portfolios, specifically in the Asian market. An earlier study by Chen and Peiser (1999) found that diversified REITs perform poorly in contrast to homogenous REIT portfolios. Moreover, Capozza and Seguin (2001) assert that homogenous REIT portfolios are simpler to monitor and are more transparent contrary to their heterogeneous counterparts. This premise reinforces the relationship between information asymmetry and heterogeneity as it can be deduced from the previous statement that diversified REIT portfolios are less transparent therefore less likely to be informationally efficient.

Cheok et al. (2011) further emphasized the notion that diversified REITs offer little marginal benefit to investors since it calls for more resources to manage properties across different regions. Additionally, they found that it would be more economically feasible to have asset managers in respective REIT firms use their knowledge and expertise of the local market to enhance the quality of the portfolio across different real estate sectors within the respective country. This implores the need for further research on the effects of diversification on REIT portfolios as well as portfolios that incorporate REITs. Diversification is partly affected by asset type.

Diversification is highly rated by portfolio managers to be an important factor in portfolio allocation and management. This is due to the fact it has been effective in

reducing the effects of systematic risk in each portfolio as compared to a geographically homogenous portfolio. Moreover, it is expected that real estate returns behave differently in different locational settings due to the difference in the market and economic conditions unique to that specific region. This is in line with the premise of real estate valuation which accords the value of a property to the macroeconomic factors affecting it as its most basic principle. Nevertheless, some authors are of the notion that geographical diversifications should be further simplified into a geographical region and an economic region as these have obvious and different implications on real estate values.

As a result, researchers have embarked on designing models that would be used to describe how heterogeneity evolves in the REITs globally. In a financial time series, it is unlikely that the variability of the “error term” will be fixed over time (Brooks 2008). Consequently, that leads to non-stochastic behaviour. This non-stochastic behaviour is due to the non-linearity of the current value of the series relative to previous values of the “error term” (Brooks 2008). Over time, researchers have found methods of modelling and forecasting volatility, namely the GARCH family model. Volatility is a key input for many financial applications, including asset allocation, risk management and option pricing (Zhou 2016).

Using the models does not solve the problem but gives direction to finding factors that influence REITs price variation. Also, the models help to understand the emergence of heterogeneity as mentioned prior. One of the factors that influence REITs price variation is “volatility spillover effects”. Volatility spillover effects can be between international markets and between different financial and real estate assets (Nikbakht et al. 2016). The literature on volatility and financial performance of REITs is insufficient without an understanding of the transmission of spillovers from other subsectors, as well as global REITs. The real estate market also revealed a significantly increased volatility, more specifically between the latter half of 2006–2009. Dania and Dutta (2017) assert that the GARCH model is what is mainly used to test for the volatility spillover effect. The knowledge of volatility linkage between different markets and assets is crucial before any testing of whether there exists volatility spillover between markets or assets are conducted (Dania and Dutta 2017).

The 2008 subprime crisis caused a market disorder that had significant adverse costs in the United States. Specifically, the finance sector was the most affected by the crisis. Financial institutions failed, specifically the banks and the conditions of the market were unsettled. The REIT industry, on the other hand, was not safe from this financial adversity. This crisis led to harsh implications on banks and financial institutions as their capacity to fulfil credit commitments was eroded; this phenomenon was observed by Huerta et al. (2016). Nikbakht et al. (2016) further stated that after the real estate and financial crisis, studies on the volatility linkage and volatility spillover between local and international received more attention. The authors further state that the volatility spillover effect is not only restricted to certain sectors, industries, or regions, for example, but financial assets and real assets are also all possible. Koutmos and Booth (1995) observed how the announcement of bad news because of these spill overs caused by the financial crisis effecting one country would cause an effect on the other countries. Similarly, when Guirguis et al. (2007) conducted an



empirical analysis in the housing market amongst regions with a different population with respect to price transmission mechanism, the results revealed a unidirectional spill over price effect from bigger regions relative to smaller regions. This empirical analysis was conducted using a bivariate General Autoregressive Conditional Heteroskedasticity model. Also, Reyes (2001) as referenced by the authors concentrating on market value indexes demonstrated using the autoregressive model that international spill over is not the only spill over effect that exists.

Concerning the lagging effect, it is crucial to first understand what the differences that exist between real estate investment trusts and common stocks are to pinpoint the impracticality of trying to use what would only work for the other securities other than the real estate investment trusts. Lin and Lee (2012) stated that real estate investment trust futures could not be hedged by existing futures used in contracts for stocks, commodities, metal, and interest rates that trying to hedge real estate investment trusts returns by the aforementioned could be a vain exercise. In that light, they further show how by coming up with the right REIT hedging returns will also enable investors to deal with the complexities that pertain to the risk dynamics that come with the REITs.

Literature shows the underlying problem that exists with the relationship between REIT values and values of the underlying real asset to be the time-varying aspect; this is often referred to as the lagging effect. Fisher et al. (2003) attributed this problem to the inefficiency of the real estate market, stating that the lack of a central market to show timeous transactions of the whole real assets causes the lag found in REIT values. While most listed REITs are required to conduct and report on the valuation of each of the properties on their portfolios, this may only be at certain periods (property valuations are done annually, on average) to disturb the market and cause insider trading by announcing sudden firm-related information. This is due to the long-standing notion that any slight change in real estate values will be echoed more swiftly by changes in REIT values (Seiler et al. 1999). They further argued that this lagging phenomenon may be due to the high transaction costs associated with direct property investment as opposed to the lower transaction costs associated with investing in securitized real estate which makes it easy for investors to use the information to their advantage. This is a concept founded by Chan (1992) when he investigated the lead-lag relation in the stock index. Sometimes the lagging effect is caused by whether an asset is listed or held in physical form.

Seiler et al. (1999) assert that unsecuritized real estate values lag those of REITs; however, he did not state to what degree which then opens it as an unexplored area of literature. Earlier studies by Gyourko and Keim (1992) and Barkham and Geltner (1995) have also found upon research the effect of the lagging effect between real property and securitized real estate; to which they found that securitized real estate leads real property assets. The issue of the appraisal methods used for securitized real estate versus unsecuritized real estate has been brought up several times in past literature. Seiler et al. (1999) differentiated the two by pointing out that the returns on REITs are derived from transaction data and those of the real estate asset are derived from appraisal data. They make the argument that transaction data obtained from the capital market where trades are more orderly thus valuations of REITs are

conducted more frequently. Nevertheless, unsecuritized real estate valuations can only be captured on average on a yearly basis and not all properties are valued at the same each year which raises other concerns. This problem has often been countered by conducting a process known as appraisal smoothing. The effectiveness of this process has been a matter of debate in past literature. With several authors pointing out the distortion it may cause to the variation and volatility measures of the real estate index, Edelstein and Quan (2006) and Seiler et al. (1999) observed how the process of appraisal smoothing can, in fact, underestimate the measured volatility in real estate returns. The appraisal smoothing phenomenon is observed by investors who rely on information obtained from real estate performance indexes, where the movements of real estate prices are displayed. The inefficiencies of these indexes are unavoidable due to the nature of real estate valuations. This is due to the subjectivity of the real estate appraisers known as heterogeneity in appraiser behaviour; as well as the time varying aspect of real estate valuation where appraisals are conducted at any time of the year.

The lead-lag concept in real estate has been attributed to the perceived inaccuracy of the valuation of unsecuritized real estate versus the accuracy of REIT appraisals. This is due to the likely occurrence of subjectivity and bias when conducting property valuations, for instance, any two valuers are not likely to use the same set of comparable properties to value a particular property due to differing opinions on the perceived relevance that these may have on the valuation. More so, the illiquidity of real estate physical assets has been associated with the lead-lag effect stating that the fact that property is illiquid and less frequently traded than securitized real estate causes a disconnect among the two.

When testing for the lagging effect, according to Abdul et al. (2010), there is robust evidence in literature that proves that there is a strong concurrent correlation that is positive, along with the lead/lag linkages between the indirect and direct real estate markets. The paper goes on to further elaborate on how the Causality analysis, which points out that the real estate market, is led to a short term by the wider economy and negative future returns may be pointed at through the positive real estate returns in the rest of the economy. This is a good demonstration of the significance of first being able to establish the relationship between the variables and how changes in them can manipulate the real estate markets over short-run and long-run horizons.

### 3 Methodology

Discrete volatilities models first emerged back in 1982 (see Engle 1982). Initially the first model was Autoregressive Conditional Heteroskedasticity (ARCH) model. Subsequently, an extension to the ARCH model was later modelled by Tim Bollerslev, which led to Generalised Autoregressive Conditional Heteroskedasticity GARCH model (see Bollerslev 1986a, b). GARCH is a robust method that can model return variation through time in a way that allows that variation to change based on the

variable's history and even when some conditions, such as price level, have not changed. Therefore, this article adopts ARCH and GARCH family models.

The formula for the ARCH (1) is as follows:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2, \quad (1)$$

where for the conditional variance to non-negative and the model well-defined  $\omega$  must be positive and all the  $i$ 's non-negative. Ding (2011) made the conclusion that ARCH models are widely used in time series analysis due to its ability to reflect the changes in variance, past econometric models often ignored this quality. Extending the ARCH model, this article chose the Asymmetric Power Autoregressive Conditional Heteroskedasticity (APARCH) and Threshold Autoregressive Conditional Heteroskedasticity (TARCH) models which are explained subsequently.

$$\sigma_t^2 = \omega + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{k=1}^r \gamma_k \varepsilon_{t-k}^2 I_{t-k}. \quad (2)$$

The TARCH model is preferred since it considers the allowance of asymmetric shocks to volatility. The introduction of the TARCH, by Glosten et al. (1993) and Zakoian (1994), was in response to the observance of varying effects of volatility due to positive and negative shocks in the financial markets. This was after it was found by Engle and Ng (1993) that when comparing positive and negative shocks of the same size, the negative shocks result in higher volatilities. Note that Eq. (1) will be lagged all the way to 1 because lagging effect has a significant role in real estate including REITs (Bowes and Ihlanfeldt 2001). Further, Ding (2011) went on to state that the PARCH model offers the best results within the ARCH type models due to its ability to express the "fat tails, excess kurtosis and leverage effect"; which are necessary to capture the conditional volatility in a time series analysis. The power ARCH (PARCH) which was designed by Ding et al. (1993) nests several of the most popular univariate parameterizations. More specifically APGARCH (p,q) model as expressed as follows:

$$\sigma_t^\delta = \omega + \sum_{i=1}^q \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^\delta + \sum_{i=1}^p \beta_i \sigma_{t-i}^\delta. \quad (3)$$

And it reduces to the standard linear GARCH (p,q) model and GJR-GARCH model (Bollerslev 1986b). The major challenge at ARCH is that it cannot be generalised; therefore, one needs a model that can be generalised. Tim Bollerslev generalised ARCH to come up with GARCH 1986. Another challenge arises in the need to have a long lag length in the ARCH models as well as the fact that its parameters are required to be of a non-negative nature (Jorge 2004).

The formula for GARCH (1) is as follows:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad (4)$$

where  $\omega > 0$ ,  $\alpha$ ,  $\beta$  and  $\lambda \geq 0$ . And  $\omega$ -constant variance coefficient,  $\alpha$ -reaction of the volatility to market events,  $\beta$  determines the persistence in volatility,  $\lambda$  captures the leverage effect,  $\alpha + \beta$  determines the degree of convergence of the conditional volatility to the long-run average level. Further,  $\alpha + \beta = 1$  otherwise the model “explodes” and  $\sigma_{t-1}^2$  is the spot variance. Note that Eq. (1) does not account for (i) correlation coefficients of debt and equity, (ii) equity parameter, (iii) risk premium, (iv) interest rates and (v) shocks-stock markets. In their analysis of various volatility models including the GARCH, EGARCH and TGARCH models, McMillan et al. (2000) found that the GARCH model offers the best results in terms of consistently forecasting prices. This notion is further confirmed by discussions held by Bollerslev et al. (1992); they contend that the GARCH model effectively represents an extensive selection of volatility processes.

According to Bollerslev (1986a),  $\omega$  is a product of the long run variance and the parameter representing the weight, for example, gamma, alpha or beta. Furthermore, the author states that when estimating the expected volatility, then  $\alpha + \beta$  represent the persistence and the greater the value the more persistence to today’s variance than to decay to the long run variance. Bollerslev (1986a) further states that  $\alpha + \beta \neq 1$  because the two parameters represent two weights out of three. Further to the GARCH model, we chose the integrated GARCH (IGARCH) extended model and the exponential GARCH (EGARCH) which we explain subsequently.

$$\sigma_t^2 = \sum_{i=1}^P \alpha \varepsilon_{t-i}^2 + \sum_{j=1}^Q \beta \sigma_{t-j}^2. \quad (5)$$

The IGARCH Model as proposed by Engle and Bollerslev (1986) is an extension of the GARCH model which, unlike the original GARCH model, incorporates the permanent effect on volatility caused by shocks due to its infinite memory. That is, the most important characteristic of the IGARCH model is its ability to capture long memory which is said to enable the predictability of returns in the long run as well as act against the effects of the market efficiency hypothesis often disregarded by other models (Ding 2011). However, Engle and Bollerslev (1986) do point out that the disadvantage of using this model is that it generates a phenomenon known as temporal aggregation which in turn reduces the model’s credibility.

The EGARCH model was developed by Daniel Nelson in 1991. This model explicitly also allows for asymmetries. However, it allows for asymmetries in the relationship between return in contrast to the AGARCH model. And  $z_t \equiv \varepsilon_t \sigma_t^{-1}$  which denotes the standardized innovations. The EGARCH (1) model is expressed as follows:

$$\log \log(\sigma_t^2) = \omega + \alpha(|z_{t-1}| - E(|z_{t-1}|)) + \gamma z_{t-1} + \beta \log(\sigma_{t-1}^2). \quad (6)$$

Negative shocks will have a bigger impact on future volatility than positive shocks of the same magnitude. This effect is normally observed with returns on stock index, and it is referred to as a “leverage effect”, Bollerslev (1986b). Moreover, Bollerslev

(1986b) stated that the leverage effect is “the tendency for volatility to rise more following a large price fall than following a price rise of the same magnitude”.

## 4 Data

### 4.1 Population and Sampling

There are 33 listed REITs on the Johannesburg Securities Exchange (JSE), and all REITs are of equity nature. This article analyses the top ten South African REITs based on their market capitalisation because they represent at least 80% of the South African REIT index. Note that the REIT legislation only became legal in South African in May 2013. Before then, listed real estate funds were either property unit trusts (PUTs) or property loan stocks (PLSs). The convention to REIT status occurred over time. The last fund to convert to REIT status is Texton Property Fund in the second half of 2019. The weekly data is from IRESS Expert database from February 2006 to December 2018. The prices are converted to log returns for consistency with ARCH and GARCH family models. The data is stratified by dates which are as follows:

- In-Sample dates from 2006 to 2018, and it is basically estimating the parameters (Akinsomi et al. 2016).
- Out-Sampling dates from 2014 to 2018, and this is for forecasting (Akinsomi et al. 2016).

Below, this article gives an overview of listed real estate funds used for the analysis.

Consisting of 454 properties within its property portfolio in South Africa that are directly owned, Growthpoint Properties Ltd is a real investment trust that also has 57 properties in Australia, a 50% interest in the properties of the V&A Waterfront, and a 26.9% interest in AIM-listed Globalworth Real Estate Investments (GWI). The combined value of these property assets was R121.3bn as of 30 June 2018. Redefine Properties Ltd is a REIT focusing on obtaining a geographically diverse portfolio of properties, as well as a portfolio of investments in listed property securities. As of 31 August 2018, the group’s property assets were valued at R91.3bn, including a South African portfolio of 315 properties with a combined value of R68.5bn.

Resilient REIT Ltd is a REIT which invests directly and indirectly in regional malls and shopping centres in South Africa, Nigeria, and Portugal, and in locally listed and offshore property-related assets. All asset management functions are housed within the company and its subsidiaries with the day-to-day management of retail centres outsourced to the property managers, namely Broll Property Group (Pty) Ltd and JHI Retail (Pty) Ltd. As of 30 June 2018, the group’s property portfolio comprised: South Africa—35 properties with a gross lettable area of 1,096,890 square metres and valued at R21, 980 m Nigeria—4 properties with a gross lettable area of 30,427 square metres and valued at R662m Portugal—2 properties valued at R1,

884 m. Hyprop Investments Ltd is a REIT status with a portfolio of properties in major metropolitan areas across South Africa, sub-Saharan Africa, and South-eastern Europe. The group's South African property portfolio was valued at R29.7bn on 30 June 2018.

Vukile Property Fund Ltd is a REIT that holds a portfolio of direct property assets as well as strategic shareholdings in listed and unlisted REITs. A group's portfolio was valued at R19.1bn on 31 March 2018. The group's property management services are currently outsourced to Broll Property Group (Pty) Ltd, JHI Properties (Pty) Ltd, Spire Property Management (Pty) Ltd, Trafalgar Property Management (Pty) Ltd, and McCormick Property Development (Pty) Ltd. SA Corporate Real Estate Ltd comprises retail, industrial, and office properties in its portfolio covering a total of 1.5 million square metres in GLA; primarily in the major metropolitan centres of South Africa, with a secondary node in Zambia. The company was valued at R16.8bn as of 31 December 2017 with a total portfolio of 196 properties. The property management function for the traditional portfolio is outsourced to Broll Property Group (Pty) Ltd, while the property management for the AFHCO portfolio is performed in-house either directly or through AFHCO Property Management.

Emira Property Fund Ltd has a property portfolio valued at R12.5bn consisting of predominantly South African assets with a growing component of offshore assets in Australia and the USA. Its sectoral profile of about 104 properties is spread across office, retail, industrial and a recent residential component. Property management services are outsourced to Eris Property Group (Pty) Ltd, Broll Property Group (Pty) Ltd and Feenstra Group. Hospitality Property Fund Ltd is a specialised REIT focusing on property investments in the hospitality and leisure sectors in South Africa. As of 31 March 2018, the Funds property portfolio comprised 53 hotels with 9,001 rooms and was independently valued at a fair market value of R12.6bn.

Octodec Investments Ltd is focused primarily on a portfolio of 306 properties located in Tshwane CBD and Gauteng. As of 31 August 2018, the group's property portfolio includes properties owned in joint ventures at a total portfolio value of R12.9bn. Octodec has an asset and property management services agreement with City Property Administration (Pty) Ltd. Fairvest Property Holdings Ltd focuses on non-metropolitan and rural shopping centres, as well as convenience and community shopping centres across South Africa. As of 30 June 2018, Fairvest's property portfolio comprises 44 properties with a GLA of 237,965 square metres and valued at R2.99bn. Property management services are outsourced to JHI Properties, Broll Property Group, Axis Property Fund, Spire Property Management, Mainstream Group, Bara Property Management and Abreal Property Management.

## 4.2 *Descriptive Data*

Table 1a, b exhibit each company's publicly preliminary and financial information.

Table 1a, b represent preliminary information of ten chosen companies. In this table are the variables of interest which are as follows: market cap, share price, net

**Table 1 a** Preliminary information. **b** Preliminary information

REIT	Market cap (ZAR Billions)	Share price (ZAR Cents)	Net asset value/share (ZAR Cents)	Sector focus (GLA in m <sup>2</sup> )							Total
				Residential	Retail	Office	Industrial	Hotel	Specialized		
GRT	84.4	2841	2843.03	0	1,390,878	1,791,626	2,254,812	0	0	5,437,316	
RDF	66.3	1159	1004.05	0	1,334,433	1,096,941	1,786,642	0	28,699	4,246,715	
RES	21.3	5000	6180.38	0	1,096,890	0	0	0	0	1,096,890	
HYP	26.9	10,822	10,487.9	0	663,505	58,956	0	0	0	722,461	
VKE	17.2	2188	2029.49	0	972,911	42,966	74,891	0	25,953	1,116,721	
SAC	12.4	488	527.08	381,907	364,801	55,241	684,478	0	0	1,486,427	
EMI	8.1	1555	1733.01	0	322,065	318,524	348,699	0	0	989,288	
HPB	6.8	1175	1931.48	0	0	0	0	9,003	0	9,003	
OCT	5.4	2040	2978.6	393,643	444,642	413,581	253,396	13,458	125,713	1,644,433	
FVT	2.1	210	239.9	0	223,216	11,748	0	0	0	234,964	
Mean	25.1	2747.8	2995.49	387,775	757,038	473,698	900,486	11,231	60,122	1,698,422	
Median	14.8	1797.5	1980.48	387,775	663,505	188,740	516,589	11,231	28,699	1,106,806	
STD Dev	27.9	3145.65	3120.09	8,299	451,608	643,551	902,323	3,150	56,820	1,753,486	
Skewness	1.6	2.2	1.9	0	0.3	1.6	0.9	0	1.7	1.5	
Kurtosis	1.4	5.4	3.4	0	-1.7	1.7	-1.3	0	0	1.5	

(continued)

Table 1 (continued)

REIT	Geographical spread (GLA in m <sup>2</sup> )											International spread	Total portfolio
	Eastern cape	Free state	Gauteng	KwaZulu Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape	Other parts of SA	Total		
GRT	163,119	0	3,371,136	598,105	0	0	0	54,373	1,196,209	54,373	5,437,315	1,234,615	6,671,930
RDF	0	0	3,207,406	293,066	0	0	0	0	680,408	0	4,180,880	0	4,180,880
RES	34,489	0	251,024	214,905	278,799	191,599	66,621	59,453	0	0	1,096,890	0	1,096,890
HYP	0	0	431,254	0	0	0	0	0	0	0	431,254	0	431,254
VKE	24,312	56,728	332,263	145,872	64,832	40,520	0	40,520	48,624	0	753,670	237,805	991,475
SAC	0	0	944,705	433,635	15,487	0	0	0	77,435	0	1,471,262	62,269	1,533,531
EMI	0	32,647	729,105	91,014	0	0	0	0	136,522	0	989,288	0	989,288
HPB	0	0	3,018	0	0	0	0	0	2,381	3,604	9,003	0	9,003
OCT	0	0	1,644,433	0	0	0	0	0	0	0	1,644,433	0	1,644,433
FVT	16,172	26,626	59,499	56,700	11,483	4,692	17,462	0	42,330	0	234,964	0	234,964
Mean	59,523	38,667	1,097,384	261,900	92,650	78,937	42,042	51,449	311,987	28,989	1,624,896	511,563	1,778,365
Median	29,400	32,647	580,180	214,905	40,159	40,520	42,042	54,373	77,435	28,989	1,043,089	237,805	1,044,182
STD Dev	69,469	15,928	1,251,694	195,992	126,448	99,199	34,761	9,799	454,847	35,899	1,779,551	632,302	2,077,141
Skewness	1.9	1.5	1.2	0.9	1.8	1.5	0	0	1.6	0	1.5	1.6	1.8
Kurtosis	0	0	0.1	-0.2	3.2	0	0	0	1.7	0	1.5	0	3



asset value per share (NAV, hereafter), sectoral focus, geographical and international spread. Table 1a shows the corresponding market capitalization amounts for the top ten REITs which ranges from a value of R2,112 billion of Fairvest Property Holdings Limited to the value of R84,406 billion of Growthpoint Properties Limited; the average, in terms of market capitalization is R25,085 Bn. According to a paper by Muller (1998) which classifies REIT sizes according to their market capitalization values in US dollars, from the data on Table 1a the top 5 REITs are either large-caps \$1.1–\$4 billion and mega-caps are +\$4 billion. And between \$0 and \$0.5 billion are small-caps and \$0.501–\$1 billion are mid-caps.

Further, Table 1a indicates that the mean NAV is higher than the mean share price, this is an interesting observation because REITs usually trade at a premium to the NAV (where share price > NAV). Therefore, we can say that the top 10 SA REIT share price is trading at a discount to the NAV. Clayton and MacKinnon (2000) argued that discounts from the NAV in REITs could be because of mispricing REIT shares due to pessimistic investor sentiment, or otherwise a result of the correct use of information on the underlying property market (information efficiency). Also on the table is the sectorial spread of the difference; one finds that most properties in the combined portfolio are retail properties, followed by industrial and then office. With regards to the geographical spread on Table 1b, one distinguishes that many of the properties are in Gauteng, followed by the Western Cape and then the Kwazulu-Natal, which is unsurprising as these provinces house South Africa's largest property markets in terms of value. It is also interesting to note that a significant portion of the portfolio of properties is located internationally, which aids for the company's diversification strategies.

Table 2 represents each company's financial information for the year ended 2018. The following are the variables of interest: revenue; total cost; profit after tax and owners' equity. The revenue is recorded in South African currency (ZAR). The same descriptive statistics measures in Table 2 will be applied to analyse the information of each company with the mean being assumed as the benchmark, standard deviation representing the risk associated within the portfolio, the skewness and kurtosis representing represent the distribution of the variables of interest representing the magnitude of the tails of the distribution, respectively. The skewness as reflected in Table 2 for both the revenue and the total cost, which is an indication of a lack of symmetry, is positive which also indicates that the possibility of diversification for both variables is paying off.

### ***4.3 Synthesis and Conclusion***

The preliminary and financial variables of interest from the ten companies analysed reveals positive results. Most companies have a portfolio mix of traditional commercial property types namely, retail, office and industrial. Amongst all, only Hospitality Property Fund specializes in hotels. However, it outperformed companies that have

**Table 2** Financial information

REIT	Revenue (ZAR millions)	Total cost (ZAR thousands)	Profit after tax (ZAR millions)	Owners' equity (ZAR millions)
GRT	10,9	182	7,9	83,2
RDF	8,1	146,5	6,6	58,1
RES	2,7	46,7	-3,3	22,9
HYP	3,1	61,9	2,6	26,4
VKE	2,0	45,1	2,4	15,9
SAC	2,3	50,2	0,8	12,9
EMI	1,8	45,9	0,8	9,0
HPB	0,9	13,4	0,1	11,1
OCT	1,9	18,4	0,5	7,8
FVT	0,4	9,6	0,3	2,4
<b>Mean</b>	3,4	62,0	1,9	25
<b>Median</b>	2,2	46,3	0,8	14
<b>STD Dev</b>	3,4	57,2	3,3	26
<b>Skewness</b>	1,7	1,5	0,7	1,7
<b>Kurtosis</b>	2,0	1,3	0,6	2,2

a diversified portfolio. As a result, this shows that the variables of interest are independent of each other. For example, low revenue will not result in a low share price as evidenced in Table 2 when comparing Hospitality Property Fund, South Africa Corporate Real Estate Ltd and Fairvest Property Holdings Ltd. Also, the size of the gross lettable area does not determine the amount of revenue the company will yield as evidenced under Table 2.

## 5 Data Analysis

### 5.1 Data Analysis

Table 3 is on historical volatilities, and it illustrates that the calculated mean is positive for most of the REITs, both for the in-sample and out-sample data. This suggests that the companies are moving with the market and that there are no benefits in diversification in most of them. Hospitality Property Fund in its nature is unique, considering that it offers organic growth, and has diversification benefits, hence the negative mean (Sebehela et al. 2019). Resilient and SA Corporate Real Estate Ltd also have negative means.

The minimum volatilities for almost all the companies are negative and positive for nearly all the maximum ones. This reflects net zero effect on the companies.

**Table 3** Historical volatilities

Parameter	Growth point	Emira	Fairvest	Hospitality property fund	Hyprop	Octodec	Redefine	Resilient	SA Corporate Real Estate Ltd	Vukile
<b>Panel A: In-Sample</b>										
Mean	0.00091	0.00042	0.00245	-0.00227	0.00136	0.00085	0.00071	-0.00133	-0.00007	0.00119
Median	0.00254	0.00218	0.00000	0.00000	0.00285	0.00000	0.00099	-0.00078	0.00000	0.00188
Min	-0.16569	-0.19479	-0.21441	-0.40213	-0.19481	-0.49470	-0.20045	-0.12464	-0.13618	-0.21806
Max	0.14552	0.16012	0.23639	0.26826	0.10764	0.51282	0.14848	0.11048	0.13976	0.13239
Std Dev	0.02990	0.03158	0.05373	0.05103	0.03082	0.04572	0.03157	0.02097	0.03284	0.03015
Kurt	2.64129	5.05598	3.31964	9.17666	3.77212	43.73591	4.00090	7.49110	2.62721	5.46518
Skew	-0.24160	-0.34457	0.39739	-0.70654	-0.54098	0.21355	-0.25108	-0.64244	-0.29579	-0.65916
<b>Panel B: Out-Sample</b>										
Mean	-0.00014	0.00010	0.00151	-0.00255	0.00029	-0.00041	0.00001	-0.00365	-0.00043	0.00055
Median	-0.00073	0.00344	0.00000	0.00000	0.00143	-0.00166	-0.00192	-0.00258	0.00000	0.00276
Min	-0.16569	-0.19479	-0.14310	-0.40213	-0.19481	-0.13238	-0.20045	-0.12464	-0.12883	-0.21806
Max	0.11840	0.11545	0.14603	0.26826	0.10088	0.09087	0.13231	0.11048	0.13976	0.13239
Std Dev	0.02881	0.03193	0.03821	0.05759	0.03299	0.03618	0.03078	0.02939	0.02783	0.02963
Kurt	4.42332	8.12851	1.39615	11.06051	5.20299	0.95767	7.37711	3.66016	4.19044	12.26624
Skew	-0.52671	-0.95146	0.35225	-0.93361	-0.85659	-0.17590	-0.76250	-0.40402	-0.03180	-1.38521

Practically, all the companies are skewed to the left, for both in-sample and out-sample. This explains that these companies can be used as investment vehicles and that a minimization of risk can occur by purely investing in them. Focusing on the skewness of in-sampling data, the overall skewness of numbers is low, meaning that risk is inherently low, and that risk can also be minimized. The out-sample numbers are more on the fringe than those of the in-sample data. The overall conclusion for the historical volatilities data from Table 3 is that there is evidence of investment conundrum—some parameters suggest investing in them while other parameters do not support investing in those REITs. However, to further continue with the getting to the objective of this research, discrete volatilities models will be used to run the in-sample and the out-sample data of the REITs (Tables 4 and 5).

All parameters of the six models are statically significant. The salient point is parameters decrease as one starts to lag them. Thus, parameters are found to be convex in shape in terms of their distribution with time. This illustrates that volatilities decrease with time, which means that risk decreases over time. This could be possibly due to diversification benefits that by nature REITs have or also perchance be due to having companies being in one portfolio as a hedging strategy. The adjusted  $R^2$  of the six models for the REITs is negative. This is acceptable in the case of real estate because there is a lot of heterogeneity in it.

ARCH (1) model well captured the market reaction of the REITs. Durbin Watson below 1.3 reflects autocorrelation. Akaike info, Schwarz and Hannan–Quinn criterion all fall outside of the range (1.6–2.7), which illustrates that there is negative skewness for all the REITs. This means that extra money can be generated from all these REITs. TARCH (1) model shows a change in delta which is an illustration of a change in spot price. The change in delta decreases with time, which is an indication of a decrease in risk over time with changes in the market. All alphas ( $\alpha$ ) and betas ( $\beta$ ) summed up for each REIT, equal to one. It shows that this model is fine. The adjusted  $R^2$ , Durbin–Watson stat, and Akaike info, Schwarz and Hannan–Quinn criterion all show similar results of ARCH (1) model. This is also seen in other models—panel C to panel F. Moreover, panel C to panel F illustrate that same patterns as ones in panels A and B. All the models confirm the same thing, which is that risk decreases over time.

All the parameters in all the six models are statistically significant. Volatilities decrease over time, which illustrate that risk decreases with time. The results obtained from out-sample show the same pattern as in the in-sample. Discrete volatilities are appropriate for risk management and risk hedging (Carr and Wu 2014).

## 6 Conclusion

The findings from this article are as follows. REITs by nature have information asymmetry, heterogeneity and lagging effects. Those three traits are confirmed by both literature review and the data analysis. Therefore, firstly, systematic risk is inherently higher in REITs. Secondly, diagnostic tests confirmed that REIT behaviour

**Table 4** Discrete volatilities: In-Sample

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
<b>Panel A: ARCH (1)</b>										
$\omega$	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***
$\alpha$	0.895 (0.000***)	0.858 (0.000)***	0.746 (0.000)***	0.832 (0.000)***	0.867 (0.000)***	0.697 (0.000)***	0.850 (0.000)***	0.918 (0.000)***	0.868 (0.000)***	0.933 (0.000)***
Adj $R^2$	-0.005	-0.002	0.000	-0.005	-0.003	-0.001	0.000	-0.013	-0.001	-0.010
<b>Durbin-Watson</b>	0.452	0.432	0.500	0.416	0.432	0.047	0.449	0.334	0.468	0.442
<b>Akaike</b>	-4.850	-4.829	-3.854	-3.992	-4.863	-4.217	-4.766	-5.894	-4.686	-4.952
<b>Schwarz</b>	-4.850	-4.823	-3.849	-3.986	-4.858	-4.211	-4.760	-5.888	-4.681	-4.946
<b>Hannan</b>	-4.850	-4.827	-3.852	-3.990	-4.861	-4.215	-4.764	-5.892	-4.685	-4.950
<b>Panel B: TARCH (1)</b>										
$\omega$	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.001 (0.000)***	0.000 (0.000)***	0.002 (0.000)***	0.000 (0.000)***	4.75E-05 (0.000)***	0.000 (0.000)***	0.000 (0.000)***
$\alpha$	0.799 (0.000)***	0.886 (0.000)***	0.667 (0.000)***	0.854 (0.000)***	0.851 (0.000)***	0.339 (0.000)***	0.642 (0.000)***	0.802 (0.000)***	0.896 (0.000)***	0.804 (0.000)***
$\beta$	0.171 (0.0523)**	-0.012 (0.886)	0.100 (0.888)	0.128 (0.087)**	0.044 (0.651)	0.058 (0.612)	0.159 (0.060)**	0.237 (0.004)***	-0.022 (0.647)	0.180 (0.035)***
$\gamma$	-0.007 (0.226)	-0.021 (0.000)***	-0.012 (0.000)***	-0.016 (0.000)***	-0.013 (0.271)	-0.017 (0.000)***	-0.033 (0.000)***	-0.005 (0.129)	-0.025 (0.000)***	-0.014 (0.000)***
Adj $R^2$	-0.004	-0.002	-0.002	-0.016	-0.003	-0.001	0.000	-0.013	0.000	-0.006
<b>Durbin-Watson</b>	0.452	0.432	0.499	0.412	0.432	0.5	0.449	0.334	0.468	0.443
<b>Akaike</b>	-4.853	-4.832	-3.844	-3.944	-4.865	-3.898	-4.754	-5.897	-4.689	-4.955
<b>Schwarz</b>	-4.843	-4.823	-3.835	-3.935	-4.856	-3.889	-4.745	-5.888	-4.680	-4.945

(continued)

Table 4 (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
<b>Hannan</b>	-4.849	-4.829	-3.841	-3.941	-4.862	-3.895	-4.751	-5.894	-4.686	-4.951
<b>Panel C: APARCH (1)</b>										
$\omega$	0.010 (0.000)***	0.011 (0.000)***	0.024 (0.000)***	0.018 (0.000)***	0.010 (0.000)***	0.019 (0.000)***	0.012 (0.000)***	0.006 (0.000)***	0.012 (0.000)***	0.100 (0.000)***
$\alpha$	0.781 (0.000)***	0.824 (0.000)***	0.690 (0.000)***	0.800 (0.000)***	0.786 (0.000)***	0.595 (0.000)***	0.786 (0.000)***	0.819 (0.000)***	0.801 (0.000)***	0.812 (0.000)***
$\beta$	0.062 (0.011)***	0.010 (0.700)	0.024 (0.367)	0.066 (0.003)***	0.020 (0.438)	0.029 (0.474)	0.001 (0.978)	0.077 (0.001)***	0.012 (0.617)	0.051 (0.040)***
$\gamma$	-0.097 (0.000)***	-0.188 (0.000)***	-0.151 (0.000)***	-0.113 (0.000)***	-0.122 (0.000)***	-0.062 (0.000)***	-0.174 (0.000)***	-0.083 (0.000)***	-0.172 (0.000)***	-0.121 (0.000)***
Adj $R^2$	-0.00487	-0.002121	-0.002869	-0.012055	-0.000665	-0.00482	-0.000556	-0.020789	-0.000323	-0.005182
<b>Durbin-Watson</b>	0.451995	0.431891	0.498324	0.4133	0.432934	0.468321	0.448549	0.330973	0.468128	0.443774
<b>Akaike</b>	-4.860169	-4.830839	-3.830447	-4.004408	-4.861588	-4.174667	-4.759218	-5.899227	-4.684325	-4.953432
<b>Schwarz</b>	-4.85104	-4.82171	-3.821318	-3.995279	-4.852459	-4.165538	-4.750089	-5.890097	-4.675195	-4.944302
<b>Hannan</b>	-4.856904	-4.827574	-3.827182	-4.001143	-4.858323	-4.171402	-4.755953	-5.895961	-4.681059	-4.950166
<b>Panel D: GARCH (1)</b>										
$\omega$	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.001 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***
$\alpha$	1.004 (0.000)***	0.980 (0.000)***	0.863 (0.000)***	0.856 (0.000)***	0.845 (0.000)***	0.425 (0.000)***	0.903 (0.000)***	0.923 (0.000)***	0.888 (0.000)***	0.947 (0.000)***
$\beta$	-0.006 (0.101)	-0.021 (0.000)***	-0.008 (0.000)***	-0.009 (0.000)***	-0.016 (0.160)	-0.010 (0.000)***	-0.022 (0.000)***	-0.005 (0.053)**	-0.024 (0.000)***	-0.013 (0.000)***
Adj $R^2$	-0.001	-0.002	-0.001	-0.005	-0.002	0.000	0.000	-0.013	0.000	-0.010

(continued)

Table 4 (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
<b>Durbin-Watson</b>	0.454	0.432	0.499	0.416	0.432	0.471	0.449	0.334	0.468	0.442
<b>Akaike</b>	-4.841	-4.833	-3.815	-3.993	-4.863	-4.014	-4.771	-5.894	-4.691	-4.956
<b>Schwarz</b>	-4.834	-4.826	-3.808	-3.986	-4.856	-4.007	-4.764	-5.887	-4.683	-4.949
<b>Hannan</b>	-4.839	-4.830	-3.812	-3.991	-4.860	-4.011	-4.769	-5.892	-4.688	-4.953
<b>Panel E: EGARCH (I)</b>										
$\omega$	-4.339 (0.000)***	-4.890 (0.000)***	-4.159 (0.000)***	-2.760 (0.000)***	-4.453 (0.000)***	-6.205 (0.000)***	-4.613 (0.000)***	-4.204 (0.000)***	-4.500 (0.000)***	-4.493 (0.000)***
$\alpha$	1.052 (0.000)***	1.099 (0.000)***	0.902 (0.000)***	0.797 (0.000)***	1.066 (0.000)***	0.893 (0.000)***	1.074 (0.000)***	1.036 (0.000)***	1.034 (0.000)***	1.111 (0.000)***
$\beta$	-0.048 (0.076)*	-0.019 (0.542)	0.006 (0.863)	-0.058 (0.001)***	-0.035 (0.236)	0.100 (0.403)	-0.006 (0.852)	-0.098 (0.000)***	-0.018 (0.530)	-0.082 (0.003)***
$\gamma$	0.537 (0.000)***	0.486 (0.000)***	0.453 (0.000)***	0.673 (0.000)***	0.527 (0.000)***	0.195 (0.000)***	0.501 (0.000)***	0.607 (0.000)***	0.504 (0.000)***	0.531 (0.000)***
Adj $R^2$	-0.005	-0.002	-0.006	-0.012	0.000	-0.005	-0.002	-0.020	-0.003	-0.002
<b>Durbin-Watson</b>	0.452	0.432	0.497	0.413	0.433	0.468	0.448	0.331	0.467	0.445
<b>Akaike</b>	-4.770	-4.742	-3.721	-3.918	-4.776	-4.078	-4.673	-5.802	-4.600	-4.857
<b>Schwarz</b>	-4.761	-4.733	-3.711	-3.909	-4.767	-4.069	-4.664	-5.793	-4.591	-4.848
<b>Hannan</b>	-4.767	-4.739	-3.717	-3.915	-4.773	-4.074	-4.670	-5.799	-4.597	-4.854
<b>Panel F: AGARCH (I)</b>										
$\omega$	0.156 (0.000)***	0.085 (0.000)***	0.152 (0.000)***	0.102 (0.000)***	0.131 (0.000)***	0.102 (0.000)***	0.133 (0.000)***	0.096 (0.000)***	0.131 (0.000)***	0.153 (0.000)***
$\alpha$	0.844 (0.000)***	0.915 (0.000)***	0.848 (0.000)***	0.898 (0.000)***	0.8699 (0.000)***	0.898 (0.000)***	0.867 (0.000)***	0.904 (0.000)***	0.869 (0.000)***	0.847 (0.000)***

(continued)

**Table 4** (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
Adj $R^2$	0.000	-0.014	0.000	-0.031	-0.003	0.000	-0.008	-0.010	-0.008	0.000
<b>Durbin-Watson</b>	0.454	0.493	0.433	0.406	0.432	0.471	0.445	0.335	0.465	0.446
<b>Akaike</b>	-4.499	-3.378	-4.495	-3.624	-4.535	-3.955	-4.431	-5.560	-4.394	-4.569
<b>Schwarz</b>	-4.496	-3.374	-4.492	-3.621	-4.531	-3.951	-4.427	-5.556	-4.390	-4.566
<b>Hannan</b>	-4.498	-3.376	-4.494	-3.623	-4.533	-3.953	-4.430	-5.559	-4.392	-4.568

Note that for every model, the first value is a coefficient and the value in the brackets is a p-value. \*, \*\*, and \*\*\* represent 0%, 5% and 10% significance level, respectively



**Table 5** Discrete volatilities: Out-Sample

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
<b>Panel A: ARCH (1)</b>										
$\omega$	0.003 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.001 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***
$\alpha$	0.899 (0.000)***	0.863 (0.000)***	0.812 (0.000)***	0.8000 (0.000)***	0.832 (0.000)***	0.793 (0.000)***	0.926 (0.000)***	0.888 (0.000)***	0.723 (0.000)***	0.935 (0.000)***
Adj $R^2$	-0.006	0	0	-0.002	-0.002	-0.001	-0.003	-0.011	0	-0.016
Durbin-Watson	0.457	0.443	0.476	0.438	0.435	0.409	0.0452	0.307	0.467	0.455
Akaike	-4.874	-4.888	-4.311	-3.75	-4.724	-4.315	-4.886	-5.165	-4.91	-5.044
Schwarz	-4.862	-4.877	-4.299	-3.738	-4.712	-4.303	-4.874	-5.153	-4.898	-5.032
Hannan	-4.869	-4.884	-4.307	-3.745	-4.72	-4.31	-4.882	-5.161	-4.905	-5.04
<b>Panel B: TARCH (1)</b>										
$\omega$	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.001 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***	0.000 (0.000)***
$\alpha$	0.823 (0.000)***	0.924 (0.000)***	0.662 (0.000)***	0.807 (0.000)***	0.788 (0.000)***	0.827 (0.000)***	0.668 (0.000)***	0.769 (0.000)***	0.747 (0.000)***	0.846 (0.000)***
$\beta$	0.216 (0.175)	-0.126 (0.320)	0.162 (0.439)	0.007 (0.963)	0.072 (0.671)	0.022 (0.903)	0.171 (0.0740)*	0.156 (0.284)	0.017 (0.934)	0.201 (0.106)
$\gamma$	-0.026 (0.000)***	-0.034 (0.000)***	-0.023 (0.314)	-0.011 (0.382)	-0.019 (0.000)***	-0.054 (0.000)***	-0.029 (0.143)	-0.012 (0.392)	-0.027 (0.001)***	-0.015 (0.002)***
Adj $R^2$	-0.005	0.000	0.000	-0.002	-0.003	-0.001	-0.001	-0.007	0.000	-0.014
Durbin-Watson	0.458	0.443	0.476	0.438	0.435	0.409	0.453	0.309	0.467	0.456
Akaike	-4.878	-4.885	-4.290	-3.748	-4.723	-4.322	-4.877	-5.164	-4.910	-5.047
Schwarz	-4.859	-4.865	-4.270	-3.728	-4.703	-4.302	-4.857	-5.144	-4.890	-5.027

(continued)

Table 5 (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
Hannan	-4.871	-4.878	-4.283	-3.741	-4.716	-4.315	-4.870	-5.156	-4.903	-5.039
<b>Panel C: APARCH (1)</b>										
$\omega$	0.010 (0.000)	0.011 (0.000)	0.016 (0.000)	0.021 (0.000)	0.011 (0.000)	0.014 (0.000)	0.008 (0.000)	0.010 (0.000)	0.013 (0.000)	0.009 (0.000)
$\alpha$	0.857 (0.000)	0.831 (0.000)	0.740 (0.000)	0.823 (0.000)	0.793 (0.000)	0.729 (0.000)	0.8781 (0.000)	0.764 (0.000)	0.726 (0.000)	0.852 (0.000)
$\beta$	0.069 (0.116)	-0.002 (0.940)	-0.016 (0.755)	0.073 (0.096)	-0.001 (0.978)	-0.028 (0.458)	0.001 (0.981)	0.052 (0.211)	0.036 (0.534)	0.074 (0.025)
$\gamma$	-0.171 (0.000)	-0.195 (0.000)	-0.155 (0.000)	-0.130 (0.000)	-0.131 (0.099)	-0.247 (0.000)	-0.130 (0.000)	-0.122 (0.000)	-0.196 (0.000)	-0.155 (0.000)
Adj $R^2$	-0.002	0.000	-0.002	-0.012	0.000	0.000	-0.002	-0.008	-0.010	-0.001
Durbin-Watson	0.459	0.443	0.476	0.434	0.436	0.409	0.453	0.308	0.462	0.462
Akaike	-4.884	-4.893	-4.321	-3.756	-4.728	-4.255	-4.914	-5.140	-4.892	-5.035
Schwarz	-4.864	-4.873	-4.302	-3.736	-4.708	-4.235	-4.894	-5.120	-4.872	-5.015
Hannan	-4.877	-4.885	-4.314	-3.749	-4.720	-4.248	-4.907	-5.132	-4.884	-5.028
<b>Panel D: GARCH (1)</b>										
$\omega$	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)
$\alpha$	0.936 (0.000)	0.894 (0.000)	0.650 (0.000)	0.812 (0.000)	0.663 (0.000)	0.841 (0.000)	0.917 (0.000)	0.887 (0.000)	0.755 (0.000)	0.958 (0.000)
$\beta$	-0.024 (0.000)	-0.032 (0.000)	-0.030 (0.000)	-0.0106 (0.363)	-0.031 (0.000)	-0.052 (0.000)	-0.025 (0.002)	-0.012 (0.171)	-0.027 (0.001)	-0.014 (0.006)
Adj $R^2$	-0.005	0.000	0.000	-0.002	-0.003	-0.0007	-0.003	-0.009	0.000	-0.015

(continued)

Table 5 (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
Durbin-Watson	0.457	0.443	0.476	0.438	0.435	0.409	0.452	0.308	0.467	0.455
Akaike	-4.878	-4.899	-4.289	-3.750	-4.710	-4.324	-4.892	-5.167	-4.912	-5.046
Schwarz	-4.861	-4.883	-4.273	-3.734	-4.694	-4.308	-4.876	-5.151	-4.896	-5.030
Hannan	-4.871	-4.893	-4.283	-3.744	-4.704	-4.318	-4.886	-5.161	-4.906	-5.040
<b>Panel E: EGARCH (1)</b>										
$\omega$	-4.019 (0.000)***	-4.817 (0.000)***	-4.988 (0.000)***	-3.548 (0.000)***	-4.410 (0.000)***	-4.635 (0.000)***	-4.439 (0.000)***	-4.408 (0.000)***	-5.026 (0.000)***	-4.633 (0.000)***
$\alpha$	1.013 (0.000)***	1.133 (0.000)***	1.030 (0.000)***	0.866 (0.000)***	1.057 (0.000)***	1.063 (0.000)***	1.259 (0.000)***	0.975 (0.000)***	0.974 (0.000)***	1.100 (0.000)***
$\beta$	-0.065 (0.173)	-0.023 (0.679)	-0.009 (0.892)	0.002 (0.958)	-0.021 (0.688)	0.021 (0.718)	0.010 (0.861)	-0.075 (0.075)*	0.025 (0.651)	-0.106 (0.010)***
$\gamma$	0.578 (0.000)***	0.492 (0.000)***	0.402 (0.000)***	0.542 (0.000)***	0.523 (0.000)***	0.467 (0.000)***	0.557 (0.000)***	0.531 (0.000)***	0.446 (0.000)***	0.516 (0.000)***
Adj $R^2$	-0.002	-0.005	-0.003	-0.009	-0.004	-0.009	-0.001	-0.008	-0.009	-0.002
Durbin-Watson	0.459	0.441	0.475	0.435	0.435	0.406	0.453	0.308	0.463	0.461
Akaike	-4.792	-4.807	-4.218	-3.608	-4.644	-4.253	-4.826	-5.032	-4.825	-4.936
Schwarz	-4.772	-4.787	-4.198	-3.589	-4.624	-4.233	-4.806	-5.012	-4.805	-4.916
Hannan	-4.84	-4.8	-4.21	-3.610	-4.636	-4.246	-4.819	-5.024	-4.818	-4.929
<b>Panel F: IGARCH (1)</b>										
$\omega$	0.2170 (0.000)***	0.180 (0.000)***	0.116 (0.000)***	0.049 (0.000)***	0.123 (0.000)***	0.142 (0.000)***	0.194 (0.000)***	0.078 (0.000)***	0.141 (0.000)***	0.138 (0.000)***
$\alpha$	0.783 (0.000)***	0.820 (0.000)***	0.884 (0.000)***	0.951 (0.000)***	0.877 (0.000)***	0.858 (0.000)***	0.806 (0.000)***	0.922 (0.000)***	0.859 (0.000)***	0.861 (0.000)***

(continued)

**Table 5** (continued)

Parameter	Emira	GrowthPoint	Fairvest	Hospitality	Hyprop	Octodec	Redefine	Resilient	SA Corp	Vukile
Adj $R^2$	0.000	0.000	-0.016	-0.004	-0.002	-0.007	-0.014	0	-0.033	-0.001
Durbin-Watson	0.46	0.443	0.469	437	0.435	0.406	0.447	0.311	0.452	-0.462
Akaike	-4.574	-4.556	-3.879	-3464	-4.396	-3.974	-4.484	-4.749	-4.631	-4.55
Schawrz	-4.567	-4.548	-3.87	-3.456	-4.389	-3.966	-4.476	-4.741	-4.623	-4.542
Hannan	-4.571	-4.553	-3.876	-3.461	-4.393	-3.97	-4.481	-4.746	-4.628	-4.547

*Note* that for every model, the first value is a coefficient and the value in the brackets is a p-value. \*\*\*, \*\* and \* represent 0%, 5% and 10% significance level, respectively

is unique—for example, low adjusted  $R^2$  are common in listed real funds. Thirdly, in- and out-of-sample data are confirming the same results. Thus, the parameters of REITs are accurately predicted.

The implications from this article are as follows. Firstly, investors need to understand that REITs are unique products, especially in terms of their volatility risk. Intraday investors can create different volatility strategies which can earn them alpha. Finally, estimation and prediction of REITs parameters, especially volatility tends to be accurate.

**Acknowledgements** We are grateful for valuable assistance from Tumellano Sebehela. The remaining errors are our own.

## References

- Abdul HNH, Anuar H, Abdul HRA, Yahya MH (2010) Relationship and lead-lag effect between Asian real estate. In: International Conference on Science and Social Research (CSSR 2010), December 5–7th, 2010, Kuala Lumpur, Malaysia
- Abdul-Baki R (2013) Do information asymmetry proxies measure information asymmetry? Published Masters Dissertation, Concordia University, Canada
- Akinsomi O, Aye GA, Babalos V, Economou F, Gupta R (2016) Real estate returns predictability revisited: novel evidence from the US REITs market. *Empir Econ* 51(3):1165–1190
- Barkham R, Geltner D (1995) Price discovery in American and British property markets. *R Estate Econ* 23(1):21–44
- Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Econ* 31(3):307–327
- Bollerslev T, Chou RY, Kroner KF (1992) ARCH modeling in finance: a review of the theory and empirical evidence. *J Econ* 52(1–2):5–59
- Bollerslev T (1986b) Glossary to ARCH (GARCH). In: Watson M, Bollerslev T, Jeffrey R (eds) Volatility and time series econometrics essays in honor of Robert Engle. Oxford University Press, Oxford
- Bowes DR, Ihlanfeldt KR (2001) Identifying the impacts of rail transit stations on residential property values. *J Urban Econ* 50(1):1–25
- Brooks C (2008) *Introductory econometrics for finance*, 2nd edn. Cambridge University Press, New York, USA
- Capozza DR, Seguin PJ (2001) Why focus matters? *R Estate Financ* 17(4):7–15
- Carr P, Wu L (2014) Static hedging of standard options. *J Financ Economet* 12(1):3–46
- Chan K (1992) A further analysis of the lead-lag relationship between the cash market and stock index futures market. *Rev Financ Stud* 5(1):123–152
- Chaudhry M, Maheshwari S, Webb J (2004) REITs and idiosyncratic risk. *J R Estate Res* 26(2):207–222
- Chen J, Peiser R (1999) The risk and return characteristics: 1993–1997. *R Estate Financ* 16(1):61–68
- Cheng P, Roulac SE (2007) REIT characteristics and predictability. *Int R Estate Rev* 10(2):23–41
- Cheok SMC, Sing TF, Tsai IC (2011) Diversification as a value-adding strategy for Asian REITs: a myth or reality? *Int R Estate Rev* 14(2):184–207
- Chung CY, Kim H, Ryu D (2017) Foreign investor trading and information asymmetry: evidence from a leading emerging market. *Appl Econ Lett* 24(8):540–544
- Clayton J, MacKinnon GH (2000) Explaining the discount to NAV in REIT pricing: noise or information? Available at SSRN 258268

- Dania A, Dutta S (2017) Examining the dynamic linkages of performance and volatility of REIT returns. *J Wealth Manag* 19(4):104–114
- Deng Y, Hu MR, Srinivasan A (2017) Information asymmetry and organizational structure: evidence from REITs. *J R Estate Financ Econ* 55(1):32–64
- Ding D (2011) Modeling of market volatility with APARCH model. Published Masters Dissertation, Uppsala University, Sweden
- Ding Z, Granger CW, Engle RF (1993) A long memory property of stock market returns and a new model. *J Empirical Finance* 1(1):83–106
- Edelstein RH, Quan DC (2006) How does real estate smoothing bias real estate returns measurement? *J R Estate Financ Econ* 32(1):41–60
- Engle R (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50(4):987–1007
- Engle RF, Bollerslev T (1986) Modelling the persistence of conditional variances. *Economet Rev* 5(1):1–50
- Engle RF, Ng VK (1993) Measuring and testing the impact of news on volatility. *J Finance* 48(5):1749–1778
- Fisher J, Gatzlaff D, Geltner D, Haurin D (2003) Controlling for the impact of variable liquidity in commercial real estate price indices. *R Estate Econ* 31(2):269–303
- Glosten LR, Jagannathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. *J Financ* 48(5):1779–1801
- Guirguis H, Giannikos C, Garcia L (2007) Price and volatility spillovers between large and small cities: a study of the Spanish market. *J R Estate Portf Manag* 13(4):311–316
- Gyourko J, Keim DB (1992) What does the stock market tell us about real estate returns? *R Estate Econ* 20(3):457–485
- Hoesli M, Reka K (2013) Volatility spillovers, co-movements and contagion in securitised real estate markets. *J R Estate Financ Econ* 47(1):1–35
- Huerta D, Egly PV, Escobari D (2016) The liquidity crisis, investor sentiment, and REIT returns and volatility. *J R Estate Portf Manag* 22(1):47–62
- Jorge C (2004) Modelling and forecasting the volatility of the portuguese stock index PSI-20. Portuguese. *J Manage Stud* 1(XI):3–21
- Koutmos G, Booth GG (1995) Asymmetric volatility transmission in international stock markets. *J Int Money Financ* 14(6):747–762
- Lin LC, Lee ML (2012) Hedging effectiveness of REIT futures. *J Prop Invest Financ* 30(3):257–281
- McMillan D, Speight A, Apgwilym O (2000) Forecasting UK stock market volatility. *Appl Financ Econ* 10(4):435–448
- Muller GR (1998) REIT size and earnings growth: is bigger better, or a new challenge? *J R Estate Portf Manag* 4(2):149–157
- Muller KA III, Riedl EJ (2002) External monitoring of property appraisal estimates and information asymmetry. *J Account Res* 40(3):865–881
- Nikbakht E, Shahrokhi M, Spieler AC (2016) An international perspective of volatility spillover effect: the case of REITs. *Int J Bus* 21(4):283–300
- Ponta L, Carbone A (2018) Information measure for financial time series: quantifying short-term market heterogeneity. *Phys A* 510:132–144
- Reyes MG (2001) Asymmetric volatility spillover in the Tokyo stock exchange. *J Econ Financ* 25(2):206–213
- Sebehela T, Marcato G, Campani C (2019) Exchange options in the REIT industry. *Adv Invest Anal Portf Manag* 9(2019):219–254

- Seiler MJ, Webb JR, Myer NF (1999) Diversification issues in real estate investment. *J R Estate Lit* 7(2):163–179
- Sufi A (2007) Information asymmetry and financing arrangements: evidence from syndicated loans. *J Financ* 62(2):629–668
- Zakoian JM (1994) Threshold heteroskedastic models. *J Econ Dyn Control* 18(5):931–955
- Zhou J (2016) Volatility is a crucial input for many financial applications, including asset allocation, risk management. *Appl Econ* 49:2590–2605

# Have Commodity Markets Political Nature?



Avni Önder Hanedar

**Abstract** This chapter examines whether good markets have political aspects, based on unique data for daily prices in İstanbul between 1918 and 1924. To set a convincing causal estimate for the impacts of political uncertainty on good prices, we focus on political risk changes during this historical episode that was not related to confounding factors, such as economic depression. Our findings shed light on the presence of higher political risk due to the resignations of governments, leading to good price fluctuations through sudden changes in supply and trade disruptions. Based on a natural experiment relating to the end of the Ottoman Empire, our results fill the gap in the literature, which covers limited research on the positive link between political events and fluctuations in commodity market prices.

**Keywords** Good markets · Commodity prices · Political uncertainty · Natural experiment · Counterfactual analysis · Risk

## 1 Introduction

Our research aims to explore whether good markets could react to political events. This issue has been discussed extensively for the financial markets which are subject to excessive speculation and arbitrage. Since the security of commodity trade depends on the political turmoil imposing significant costs on businesses and investments, political risk is also one of the important determinants of commercial activity. This would reflect fluctuations in the prices of commercial goods, for instance, if economic agents perceive a serious threat to investments and production. Over the last decade, the volatility of commodity prices is rising over time because of serious political events such as wars and regime changes (Bittlingmayer 1998; Anderson and

---

A. Ö. Hanedar (✉)

Faculty of Political Sciences, Sakarya University, Esentepe Kampüsü Sakarya, Serdivan 54187, Turkey



Marcouiller 2002; Su et al. 2019; Hou et al. 2020). The effects of political changes on commodity prices have not received adequate attention yet. In this respect, our research question is how the real economy responds to the uncertainties created by political events. Our results could herald a further discussion of the role of the uncertainties on commodity prices.

The literature on the relationship between political uncertainty and economic outcomes addresses several mechanisms to induce commodity price fluctuations in case of serious political risk. First, political events are generally less predictable in comparison to economic ones. So this nature could generate significant levels of supply disruptions and economic hardship expectations in the future. Second, the political turmoil could lead to an environment of uncertainty about future economic policy decisions. This could create lower demand for households through increasing transaction costs. Third, higher political uncertainty provides negative signals on the future of firms and traders' life, leading to lower investments and decreasing demand for inputs (Anderson and Marcouiller 2002; Julio and Yook 2012; Baker et al. 2016; Jens 2017). Last, during the persistent political unrest and turmoil, commodity producers might cut back on their production and lay off workers, heralding lower economic growth (Asteriou and Price 2001; Hou et al. 2020).

In the last decades, regime changes and armed conflicts are more often observed. These events disrupted markets worldwide, which could be responsible for the remarkable volatility in commodity prices in the presence of higher speculation. However, there is a dearth of research for the effect of political risk on commodity price volatility (Hou et al. (2020) and Zhu et al. (2020) for discussion on economic policy uncertainty in case of political turmoil). One exception is the energy market, covering many discussions on the role of the uncertainties for the price fluctuations. The remarkable volatility of energy goods' prices has sparked a debate about the positive impacts of political risk (Yin and Han 2014; Bouoiyour et al. 2019; Shen 2020). This is because energy goods are essential materials for production and are closely related to political patterns (Su et al. 2019). The lack of research on how political uncertainty induced good price fluctuations is an important motivation of our study. We also find it relevant to understand what type of political crises could generate larger shocks of commodity markets, while riots in Africa and Asia of the 2000s due to agricultural good price fluctuations triggered new tensions (Yin and Han 2014).

Another important contribution of our chapter is its original methodology, which is based on a historical experiment allocating observations to different treatments randomly. There are increasing number of papers that are exploiting natural experiments based on historical rare events. For instance, the Great Mississippi Flood of 1927 is used to test the effects of lower agricultural labor on agricultural development (Hornbeck and Naidu 2014). Besides, the reunification of Germany is selected as a natural experiment to test the role of market access on economic development (Redding and Strum 2008). In a similar vein, our paper uses a historical case to deal

with bias because of omitted factors. These variables relating to political risk and prices, such as economic depression, might conflate the effects of the political events on prices and volatility, suggesting a  $cov(Prices, u_i/\beta_1) \neq 0$ . This means that the OLS estimate for the effects of political risks on good prices,  $\beta_1$ , might be biased. More recently, to address this issue, research has focused on a natural or quasi-experiment as a source of exogenous variation in the variable of interest such as political risk, which is not related to confounding factors. Like our study, Wang (2019) and Wang and Boatwright (2019) used 1995–1996 Taiwan Strait Crisis as a natural experiment to establish a persuasive causal link between political risks and financial markets.

To test the existence of an unbiased causal relationship between political risk and good prices, we choose an important historical period from 1918 to 1924 of the Ottoman Empire. We have three special reasons for choosing this case. First, for this period, a unique data set on the good prices in İstanbul is available. Second, the events during this period are of a political nature, whose timing was independent of local and global economic conditions.<sup>1</sup> Last, this historical period offers an opportunity to estimate a counterfactual analysis to capture the impact of political risk on the economy (Bittlingmayer 1998; Estevadeordal et al. 2003; Wang 2019). The causal effect would be derived from the difference between counterfactual outcomes assuming full control of the state on prices and actual prices under the effect of different political events.<sup>2</sup> During WWI, the Ottoman state intervened in the commodity markets with urgent and strict regulations (Eldem 1994, pp. 18–19),<sup>3</sup> which represents our lower price fluctuations benchmark. A period of higher risk could be after the ceasefire period between September and November 1918. This is because members of governments resigned at the end of 1918, and the future of the country was uncertain by the beginning of 1919. The administrative structure deteriorated, although people expected that economic progress would happen soon thanks to peace, creating a price fall for a while (Tasvir-i Efkâr 16 December 1918, p. 1). Moreover, the occupation and civil war had begun after February 1919 while two separate government authorities in different parts of the Ottoman Empire controlled the economy (Ediger 2007, p. 340). As shown in Fig. 1, the number of articles in the Ottoman newspapers on economic policy regulations—such as price ceiling—decreased by 1922. The number of articles on economic regulations reached its highest point in 1923 when the strong control of the economy was established thanks to the end of the occupation by October 1923 (Çavdar 1983, pp. 64–65).<sup>4</sup> Meanwhile, the amount of news on price instabilities increased as the occupation began in 1919. The higher number is parallel to decreasing amount of news on the

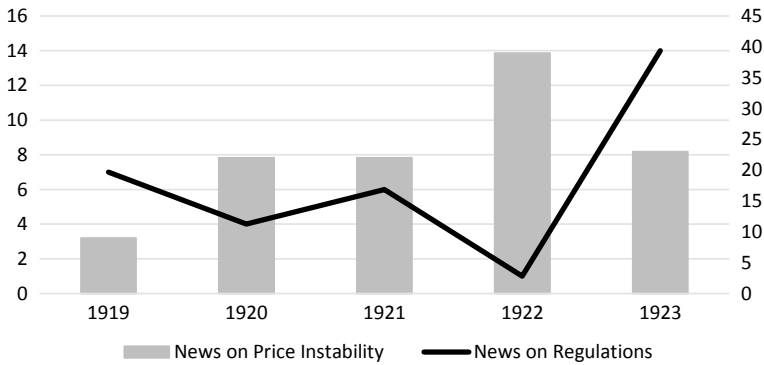
---

<sup>1</sup> Recently, Hanedar et al. (2018) find that the outbreak of many political events from 1918 to 1924 was not related to bond prices at the İstanbul bourse, supporting the independence of such events from the economic conditions.

<sup>2</sup> For instance, between 1918 and 1925, the Ottoman newspapers, such as *Vakit*, had sections of the price ceiling that was imposed by the government. The efficiency of this policy could be related to the government's power, which could be diminished by the regime changes.

<sup>3</sup> See also Toprak (1982) and Hanedar (2018) for a detailed discussion on economic policies.

<sup>4</sup> Under the lack of detailed data, for instance, the articles in the newspapers could reflect outcomes of the economic policies and fluctuations efficiently (Shiller 2017).



**Fig. 1** The number of newspaper articles on economic policy regulations and price instability in the Ottoman newspapers, 1919–1923. The data are compiled from *Vakit* and *Tasvir-i Efkar*. Point cover the number of price ceiling lists that could be daily published during the period. This is because there is not sufficient information at this stage about whether these lists were regularly published in *Vakit* during the period. The data after 1923 are not included, as the newspapers included news on the political transformation of the country

economic policy regulations. This could support our paper’s usage of a natural experiment based on the positive effect of political problems on decreasing control of the economy and price instability.

In a nutshell, our paper has two specific purposes. First, we aim to quantify the causality between commodity prices and political risk. To capture the causal effect of the political crisis, we construct a counterfactual analysis based on the dissolution of the Ottoman Empire between 1918 and 1924. The effect of the political events will be measured as the gap between the actual prices and the counterfactual path during the end of WWI when the state had relatively enough power to control prices. In other words, to determine the magnitude of the political risk leading to higher prices, our paper answers what if the political fluctuations had not happened in the Ottoman Empire after November 1918. Then we can test the presence of sufficient evidence for the political uncertainty priced. Second, our study applies ARCH methodology on volatility changes to ensure the magnitudes of the political risk premium for a group of events. We examine which political crises had a noticeable impact on commodity prices. More specifically, this exercise is helpful to predict the responsiveness of good markets to various political events. Here, it is necessary to state that marrying this historical case with an econometric perspective would reveal the importance of the study’s contribution to literature without sufficient quantitative information on this subject. Thus, we have three important findings. First, our findings shed light on the presence of higher risk perception only when the government’s survival was problematic. This result may offer a policy suggestion that the worst government is better for the economy and social order than no government case in case of political uncertainties. Second, we find that political risk generated the biggest volatility in the prices of energy goods that were imported abroad, adding the findings of the literature on political risk. Last, our study shows how researchers can use natural

experiments and counterfactual models to provide persuasive causal evidence without lab-controlled settings.

The rest of this chapter is organized as follows. Section 2 presents previous literature on the effects of political risk on commodity markets and papers about natural experiments. Section 3 presents the data and method. Section 4 discusses our empirical findings. In the last section, concluding remarks are summarized.

## 2 Literature Review

In the literature, the effects of political risk on good markets have not received attention, as there is a lot of research on the impacts of political uncertainty on bonds and stocks (Wang 2019). Good prices might be highly responsive to political events as well. To sum up, wars could disrupt trade routes and cause damage to products, meaning a high risk for traders. In case of regime changes, governments could not control the markets efficiently. Also, there could be disruptions in the economic policies and firm activities, which deteriorates the value of goods (Asteriou and Price 2001; Anderson and Marcouiller 2002; Hou et al. 2020). A pioneering paper on the relationship between political uncertainty and economic outcomes is Bittlingmayer (1998)'s study. Bittlingmayer (1998) examines the changes in the link between economic outcomes and stock market volatility during exogenous political events of the early twentieth century in Germany, such as regime changes. These events are natural experiments, that were exogenous determinants of political risk creating stock market volatility and output shocks, to eliminate the bias due to endogeneity. Bittlingmayer (1998) shows a strong and unbiased causal effect from political risk to stock market volatility and economic output shock. More recently, to deal with such problems, natural experiments are often used to create exogenous factors affecting dependent variables. Unlike lab experiments, natural experiments are based on real exogenous phenomena that can be a product of environmental, social, or political forces. For instance, like our study, Redding and Strum (2008) employed political events, such as the division and reunification of Germany, to deal with bias in econometric analysis to test the role of market access on economic development.

Different from conventional studies on the relationship between political uncertainty and financial markets, Wang (2019) and Wang and Boatwright (2019) examine the link between political risks and financial outcomes from the perspective of natural experiments. In Wang (2019) and Wang and Boatwright (2019)'s study, Taiwan Strait Crisis between 1995 and 1996 is used as a natural experiment. They construct a counterfactual model to check the causal link further. These papers find a positive effect of political risks on financial volatility during the Crisis. There is a lack of such literature for the collapse of good markets in case of political risk. One exception is Hou et al.'s (2020) working paper. Similar to Wang (2019) and Wang and Boatwright (2019)'s papers, they show that political uncertainty is statistically correlated with changes in the volatility of good prices in 12 countries. As political risk could be related to confounding factors such as economic uncertainty, an unbiased estimate is

possible after testing the uncertainty during the US presidential election, which is a natural experiment to predict the political risk. While this paper covers preliminary results, its findings are in line with a related branch of the literature focusing on the impact of political uncertainty on firm decision and growth. This literature indicates that during higher political uncertainty demand by households would shrink, which would lead to a price decrease. Also, producers could stop production, which brings lower employment and higher price (Asteriou and Price 2001; Julio and Yook 2012; Baker et al. 2016; Jens 2017).

### 3 Data, Setup, and Model

#### 3.1 Data and Setup

To unveil the effects of political changes on risk perception for the economy's future, we use prices of various goods in İstanbul from 30 May 1918 to 16 May 1924. The daily price data are elaborated from the Ottoman newspapers, *Vakit* and *Tasvir-i Efkar*. The prices are expressed in Turkish Lira. We collect various parts of the data using the list of price ceiling, which are reported in the newspapers as shown in Fig. 2. The market prices could be different from those in the lists, as the traders could charge higher prices in case of lower power of government to protect consumers and impose laws.<sup>5</sup> The data sources include daily observations of more than twenty different products. For instance, the data cover prices of gas and petroleum products used for heating and lighting, such as Romanian and Georgian gas and petroleum. In addition, there were prices of agricultural commodities such as barley, bread, olive oil, and wheat.

Price data on individual goods have missing observations. For instance, the data for gas from Romania are not available for several days. Price data for gas from Georgia have missing observations for different days. For many of the goods, at least two observations are available. We separately create price indices for different goods to get a variable with fewer missing observations by considering the earliest dated observation as base or 100. Then we take an average of all indices because of data limitation on the consumption weights. We divide them into two groups. These are energy goods and commodity price indices. The base date is 30 May 1918 for these indices. We have an important motivation to create two individual indices, which is to distinguish the effects of political risks on imported and domestically produced goods. Energy goods were imported as commodities were mostly agricultural goods, which were produced domestically. However, the negative effects of WWI on production might sometimes cause the purchase of the commodities from abroad (Eldem

---

<sup>5</sup> In the list, it was stated the presence of some penalties in case of failure to comply the prices listed. In addition, there could be a limited flexibility to set prices for some of the goods such as unlisted commodities and coal (Vakit, Azami fiyatlar, 14 June 1922, p. 4).

**اعظمی فیات**

۱۴ تموز ۳۳۸ تاریخند ۲۰ تموز ۳۳۸ تاریخند قادر منبر اولی اوزره مواد غدایه  
وحوالی ضروریه به موضوع اعظمی فیاتی میل ایسته در :

اسمی	فیات	اسمی	فیات	اسمی	فیات
مارتیسفورنجی اجنبی اونز	۲۲	جاوا مال	۴۰	شکر	۴۰
دوردم ایکنبی	۱۸	مولاندا کوب	۴۸	ایجه طوز	۱۰
اکسترا برلی	۲۰	قازه	—	ابری بملک سوغان	۸
ایکنبی	۱۷	بلجیقا کوب	۴۸	اسکندریه سوغانی	۸,۲۰
آسرقا برنجی (بلوروز)	۳۸	اکسترا اکسترا بملک زیتون یانی	۸۰	قیورجیق قویون آتی	۹۵
سیام	۲۳	برنجی	۷۵	ایکنبی قیورجیق	۸۲
اسبابا	۲۵	ایکنبی	۷۰	برنجی داغلیج قرمان	۹۵
انکیز	۱۹	اکسترا اکسترا کولیه صابونی	۳۶	ایکنبی	۸۲
برنجی قیرتلی	—	خالص طرزون تره یانی	۱۴۰	اوجنبی	۷۲
برلی اسمر مفارنه	۳۰	ایکنبی	—	خالص سود	۲۸
ایرمیک مفارنه سی	۳۴	برنجی آسرقان یانی	۷۰	ساده طمن حلواسی	—
بسی چالی فاصولیه سی	۲۰	ایکنبی	۶۸	کلاج	—
اوزنه	۱۸	اوجنبی	—	شندوفر کوری (ده بورده)	۷
طرزون سرمالی	۱۳,۲۰	ردم ایل بیاض پنیری برنجی	۱۰۸	دکانزده	۷,۲۰
خروس	۱۹	بولغار پنیری	۹۷	آناطولی کوری ده بورده	۶
رومانیا سرمالی فاصولیه سی	۱۳,۲۰	طلووم پنیری	—	محلانده	۶,۲۰
اسکندریه بتانی	۸,۲۰	برلی برنجی زیتون	۳۸	کیسلش قوری اودون دکانده	۲,۲۰
آمله یازاری کله پاتانی	۸,۲۰	ایکنبی	۳۰	یاش	۲
اونق	—	اوجنبی	۲۰	کیسله مش قوری میشه ویا کورکن	—
ایتالیا پاتانی	—	برنجی آسرقا غازیانی دومل	۱۸	اودونی (ده بورده) چکیسی	۳,۲۰
مولاندا خالص قوستالیه توز شکر	۱۴	فوشلی	۱۵	کیسله مش یاش میشه ویا کورکن	—
آسرقا توز	۴۰	دوکه رومانیا	۱۲,۲۰	اودونی (ده بورده) حکسه	۲,۸۰

Fig. 2 The price ceiling for goods in İstanbul. Point figure is taken from *Vakit* (Azami Fiyatlar, 14 June 1922, p. 4).

1994, pp. 48–50; Ediger 2007, pp. 339–340). The final sample consists of 170 daily observations for energy goods and 165 observations for consumer goods.

Table 1 The average prices during the political events between 1918 and 1924

	$P_{t-energy\ goods\ prices}$	$P_{t-commodity\ prices}$
The First World War	129	90
	(51)	(51)
The resignation of governments	204	73
	(71)	(71)
The resistance and occupation	152	106
	(35)	(30)
The Republic of Turkey	161	77
	(13)	(13)

The number of observations are reported in parentheses

Table 1 presents summary statistics on the average values of the prices in the periods during the events examined. To introduce the necessary background, we will discuss how the political uncertainty due to the different political events was reflected in the commodity market of İstanbul.

According to Table 1, when the end of WWI was approaching by November 1918, the energy goods' price is observed as 129 units and the commodity price is 90 units, on average. During WWI there were strict regulations on commodity markets due to close economy in case of war circumstances, although energy goods' prices had increased due to disruptions of trade link between the Ottoman Empire and its Allies by the end of the conflicts (Eldem 1994, pp. 18–19, 79–180; Hanedar 2018). For this period, Hanedar (2018) points out that while WWI had destructive effects on the economy, the ceasefire period created temporary economic progress and a price decrease. This is because traders expected that restrictions and damages would be lifted soon, creating an increased supply of goods.

After the end of WWI, the Allies organized several meetings to negotiate their demands from the defeated countries. The US proposed Wilsonian principles to secure the end of the conflicts and to deal with a long list of these demands, providing relative protection for the Ottoman Empire's lands. In 1919, the Allies suddenly occupied İstanbul and Greece controlled İzmir, which could have triggered resistance and civil war (Fromkin 2001, pp. 404–407). Before the Allies' occupation, the members of the war government fleet and the political instabilities become crucial (Hanedar et al. 2016). Because of the government's resignation, there was a period of institutional problems with the control of prices. During the resignation of governments after the end of 1918, the energy price is 204 units, as commodity price is 73 units, on average. Although commodity prices are slightly lower than that of WWI, the energy goods' price was twice as high. This could be related to increasing uncertainty and trade disruptions, as in December 1918 *Tasvir-i Efkâr* argued an increase in prices of many goods, such as flour, gas, soap, due to institutional weakness during the resignation of governments in comparison to the armistices' period by November 1918.<sup>6</sup> The one part of the cartoon presented in Fig. 3 was taken from the article to provide further insight on this issue. In the cartoon, the flour was presented and increases in the prices are expressed by the growth in the picture of the good. Because of the resignations and weak institutions, Fig. 3 shows the price increase in the flour after the ceasefire period.

When a new and stable government was appointed at the beginning of 1919, the occupation had already begun, which would have created price fluctuations by 1923. In 1923, the civil war ended, and a strong government was established in Ankara (Çavdar 1983, pp. 64–65). During the resistance and occupation, energy price is 152 units, while consumer good price is 106 units, on average. In this period, due to ongoing political conflicts, the prices were higher than those of WWI. Price increase could be also mentioned in the Ottoman newspapers. For instance, *Vakit* provided news that the occupation of İzmir induced prices.<sup>7</sup> This could be arisen by lower

<sup>6</sup> *Tasvir-i Efkâr* (16 December 1918) Gala-yi es'ar, p. 1.

<sup>7</sup> *Vakit* (19 May 1919) Piyasa, p. 2.



**Fig. 3** Inflation after the Armistice of Mudros. Point figure is taken from *Tasvir-i Efkar* (Gala-yi es'ar, 16 December 1918, p. 1)



trade as the import and transportation of goods from Anatolia become problematic.<sup>8</sup> So, as the resistance was ongoing in 1920 it was argued that as compared with the other belligerents the Ottoman Empire had the highest price increase. The inflation rate was about 1300%. Prices in France and UK increased by 162 and 269%.<sup>9</sup> As the Allies' forces were defeated by the Turkish forces in 1922, there were short-run improvements in economic life thanks to government control.<sup>10</sup> After the establishment of the Republic of Turkey, the prices are 161 and 77 units, respectively. There was a decrease in the commodity prices, as there could be a lack of energy goods again. Table 1 shows that during the foundation process of the Republic of Turkey, energy prices have risen. By the foundation of the Republic of Turkey in October 1923, it seems that the value of the Lira was positively affected by government interference,<sup>11</sup> leading to lower speculations. Due to this situation, the prices of wheat, bread, and flour become stable.<sup>12</sup> Overall, the descriptive findings suggest that government changes and civil war could be parallel with higher price fluctuations through increasing political risk, while the establishment of a new regime could be inadequate to cope with problems at the beginning.

<sup>8</sup> *Tasvir-i Efkar* (1 June 1919) Piyasa ahvali, p. 2.

<sup>9</sup> *Vakit* (2 October 1920) Nasıl yaşayabiliyoruz? p. 2; (4 October 1920) Piyasa, p. 3.

<sup>10</sup> *Tasvir-i Efkar* (7 August 1923) Borsada ihtikarın menine doğru, p. 3; (2 September 1923) Borsa komiseri Adli bey'in beyanati, p. 3; (4 September 1923) Galata borsasında istihale başlıyor, p. 2.

<sup>11</sup> *Vakit* (23 September 1923) İktisadi hafta, p. 3; (6 October 1923) September maaşı, p. 2.

<sup>12</sup> *Vakit* (28 September 1923) Ekmek narhı, p. 3.



### 3.2 *Econometric Model*

Similar to Wang (2019) and Wang and Boatwright (2019), our paper runs an empirical analysis, grounded by a counterfactual model to examine the effect of political risk due to exogenous events. To identify whether there exists a causal effect of political risk on prices of energy goods and commodities, our study constructs a model as follows:

$$P_t = \beta_0 + \beta_1 ER_t + \beta_2 W_i + \beta_3 T_t + u_t, \quad (1)$$

where  $P_t$  is price indices of energy goods and commodities in day  $t$ .<sup>13</sup> We also use conventional determinants of the prices (Wang et al. 2019) to control the effects of other covariates. In order to approximate the prices of complementary and substitutes, we use  $ER_t$ , which is the value of Gold against the Turkish Lira in day  $t$  at the İstanbul bourse. This variable also captures the fact that the energy goods were imported abroad, whose prices were affected by exchange rates.<sup>14</sup> In Eq. (1), as daily observation for supply and quantity traded are not available,  $W_i$  is used as a proxy for economic outcomes. This variable reflects the average precipitation reported as centimeters, in year  $i$ . The higher precipitation could be expected to be positively related to demand and supply in an agricultural economy through increasing production and income.<sup>15</sup>  $T_t$  is a time trend to capture the impacts of economic progress and consumer preferences.  $u_t$  is an error term.

We design the analysis to capture the differences in prices between periods of high and low political uncertainty. The period before the end of WWI in November 1918 was the time of lower price fluctuations due to strict government control. The other periods are regarded as the time of high political uncertainty due to the occupation and resistance, which would have created serious price changes. Equation (1) finds the counterfactual prices as if there would not be serious and observed uncertainty like during the end of WWI. The causal link from political risks to commodity prices will be inferred by posing the following counterfactual question:

What kind of a path would commodity prices have followed from 1918 to 1924 if government control remained the same as the periods of WWI?

Thus, we expect that the predicted prices would be lower than real prices if there is a causality running from exogenous political shocks to higher risk, through decreasing the government ability and leading to disruptions in the market mechanism.

---

<sup>13</sup> To compare predicted prices with observed values easily, our paper does not use the logarithmic transformation of the variables.

<sup>14</sup> The data are compiled from the Ottoman newspapers, i.e., *Vakit and Tasvir-i Efkar*.

<sup>15</sup> The data come from National Oceanic and Atmospheric Administration (NOAA) and only available at the year level. The data on weather, such as precipitation, is a good proxy for production in economies based on the agricultural sector (Hanedar 2016; Kalemli-Özcan et al. 2020).

Data constraints on the determinants of the prices limit further analysis based on the counterfactual method. Without controlling many other covariates, we could check the changes in the volatility of prices using the ARCH model. Also, this model makes us understand which events were risky. After testing the causality between political events and prices, our paper estimates an ARCH model<sup>16</sup> including separate dummy variables for exogenous political events in Eq. (1):

$$h_t = \omega + \alpha EVENT_t + \beta h_{t-1} + \varepsilon_t, \quad (2)$$

where  $h_t$  is the variance of the return of price indices for energy and commodities at day  $t$ . The study calculates returns of the good prices as follows:

$$R_t = \ln(P_t/P_{t-1}). \quad (3)$$

Equation (2) relates to a long-term average ( $\omega$ ) and information about the variance of the last period ( $h_{t-1}$ ) with volatility in the prices  $h_t$ . *EVENT* equals one for the dates of the events and zero for the period by the end of WWI. In our case, *EVENT* includes three separate dummy variables. First one takes one during the dates of the resignation of governments between November 1918 and the beginning of 1919, and zero otherwise. Second dummy variable is one for the dates of the resistance and occupation between 1919 and 1923, and zero otherwise. Third dummy variable takes one after the Republic of Turkey was established in 1923, and zero otherwise. If there was high political risk at the date of political events in our treatment group, then this implies a positive  $\alpha$ . This points out the difficulty to deal with price fluctuations. For instance, the resignation of governments could mean increasing uncertainty, which would create higher price changes due to lower investments and supply of goods. This is because consumers and investors could not predict what would be in the future. In addition, in the case of weak institutions and lower state control, the prices could be more fluctuated.

## 4 Results

Table 2 shows the coefficient estimates in Eq. (1) to predict the prices with conventional determinants of demand and supply by assuming the absence of any uncertainty to set the prices. Our result implies that some conventional variables could statistically predict the prices of energy goods and commodities. We find that depreciation of the Lira is correlated with increasing prices of energy goods. This means that energy goods' demand was negatively affected by the costs of imports as these commodities were coming from abroad. A lower value of the Lira, however, decreased the commodity prices. This finding might imply that expensive commodities such as

---

<sup>16</sup> The coefficient estimates of interests are robust to the inclusion of GARCH effects, which are not statistically significant.

**Table 2** The prices under the absence of risk between 1918 and 1924

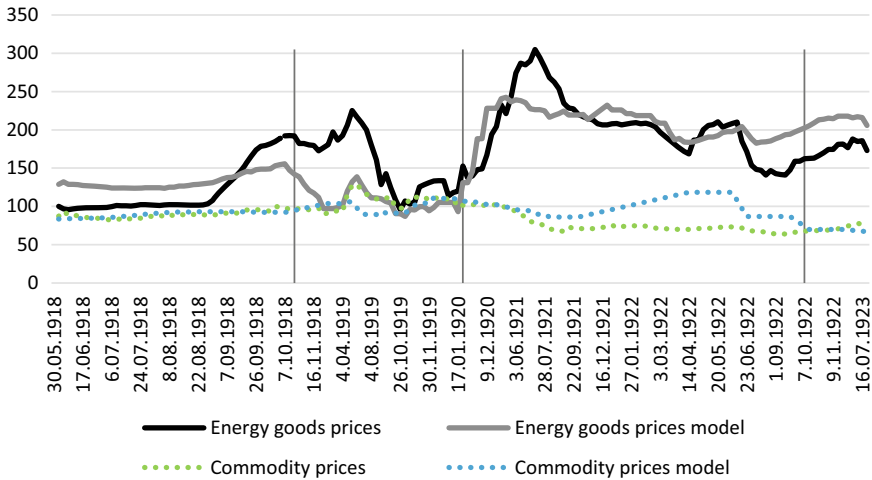
	<i>P<sub>t</sub>-energy goods prices</i>	<i>P<sub>t</sub>-commodity prices</i>
<i>ER<sub>t</sub></i>	0.356* (0.003)	-0.108** (0.025)
<i>W<sub>i</sub></i>	2.598 (0.293)	0.711*** (0.092)
<i>T<sub>t</sub></i>	-0.312 (0.148)	-0.019 (0.845)
Constant	-2.784 (0.971)	152.660* (0.000)
No. of obs.	103	103
<i>R</i> <sup>2</sup>	0.386	0.173

All equations are estimated by the OLS. Standard errors are heteroskedasticity-robust standard errors. *p*-values are reported in parentheses. \*, \*\*, and \*\*\* denote statistical significance at 1%, 5%, and 10%, respectively

wheat would be substituted for cheaper imports since the production had been permanently damaged by WWI (Çavdar 1983, pp. 64–65; Eldem 1994, pp. 47–50). Finally, higher precipitation led to an increase in commodity prices. This result could be explained by the devastation of crops in case of bad weather.<sup>17</sup>

In Fig. 4, we present observed and predicted prices based on models in Table 2. The vertical lines show the dates of the end of WWI (the first part), the resignation of the governments (the second part), the occupation and resistance (the third part), and the foundation of the Republic of Turkey (the fourth part), respectively. Predicted prices of energy goods and commodities are higher than the observed values during the end of WWI (the first part). This gap is low in commodity prices. The finding could provide evidence on the presence of relatively moderate risk. The lower fluctuation of prices is in line with Hanedar (2018)'s arguments showing that the end of the military conflicts was associated with temporary economic progress and lower prices for many goods, based on the Ottoman newspapers. During the resignation of the governments (the second part) predicted prices of energy goods and commodities are lower than the observed values. In particular for energy goods, this gap is high in favor of observed prices. These results mean that political uncertainty after the resignations of governments might create price fluctuations. When the occupation and resistance began (the third part), energy goods' prices were too high. At the same time, there was a decreasing trend in commodity prices. The predicted and the observed energy goods' prices are similar. On the other hand, in 1922 predicted prices of the

<sup>17</sup> Coefficient estimates are robust to the inclusion of additional control variables, such as the number of good and bad news mentioned by the Ottoman newspapers on the economy. In addition, the value of the Lira could reflect some political events' impacts on prices (Hanedar et al. 2019). To cope with this issue, we add separate dummy variables for political events examined in our study, which does not make large changes in the coefficients of interest. This also provides evidence supporting the exogeneity of the events.



**Fig. 4** The observed and predicted prices, 1918 and 1924. The data come from *Tasvir-i Efkar* and *Vakit* and are obtained related columns in Table 2

commodities are higher than the observed ones. This finding suggests a lower level of institutional weakness though the presence of occupation and resistance. During this period there could be a higher capacity of occupiers to control prices as well as mitigating the negative effects of WWI through increasing trade and production. In addition, a lack of demand and depression could have led to a lower increase in prices.<sup>18</sup> Finally, for energy prices, the observed values were lower than the predicted ones after the foundation of the Republic of Turkey (the fourth part). Meanwhile, the observed and predicted prices of commodities have a similar path. This situation can be regarded as a sign of lower political uncertainty thanks to the establishment of the new regime and the end of the occupation, which would lead to powerful institutions and an efficient price mechanism. To sum up, in line with our argument, we notice that by the end of WWI prices were relatively stable, as price fluctuations were observed during the resignations of governments, which could imply a higher risk for the business.

Table 3 presents the findings of the ARCH model in Eq. (2). The results indicate a statistically significant increase in volatility only during the resignation of governments. The coefficient estimates are 4.112 and 1.731, respectively. We have no evidence that people were worried about the supply of goods during the presence of a war government, occupation, and a continuing state. This finding suggests that when the governments resigned the price, fluctuations were high, reflecting higher political uncertainty. The result is in line with Fig. 4, showing the presence of a gap between observed and predicted prices, due to increasing uncertainties on the country’s future during the resignation of governments (the second part). Also during

<sup>18</sup> *Vakit* (16 June 1921) Piyasa, p. 3.

**Table 3** Risk and political events

	$h_{t-1}$ -energy goods prices	$h_{t-1}$ -commodity prices
The resignation of governments	4.112* (0.000)	1.731** (0.030)
The resistance and occupation	1.055 (0.317)	-0.256 (0.737)
The Republic of Turkey	0.476 (0.840)	1.584 (0.104)
$h_{t-1}$	0.751** (0.049)	0.570* (0.003)
Constant	0.004 (0.840)	-4.130* (0.000)
No. of obs.	169	163

Standard errors are heteroskedasticity-robust standard errors.  $P$ -values are reported in parentheses. \* and \*\* denote statistical significance at 1% and 5%, respectively. N. of obs. is the number of observations

this period, Fig. 1 shows a lower amount of the news on regulations imposed by the state, when the Ottoman newspapers often mentioned the news on price instabilities.

Our study highlights that some kinds of political events posed risks to the economic activities reflected in the good price fluctuations. We find that political risk could produce large fluctuations in the prices during the resignations of governments because of the lack of powerful institutions to mitigate the negative impacts of WWI on trade and production, as argued by *Tasvir-i Efkar* (16 December 1918, p. 1). Our result also shows that the energy and import markets are more sensitive to political instability. By providing new insight on the energy market with a historical case, our paper contributes to the literature on the effects of political risk on energy goods (Yin and Han 2014; Bouoiyour et al. 2019; Shen 2020). From the Ottoman context, we can propose that in the case of weak political institutions or regimes, it could be difficult to cope with supply and price shocks due to the serious political transformation. Surprisingly, our analysis does not indicate the sensitivity of the commodity market to all kinds of political events, such as occupation and civil war. Economic progress under the control of occupiers might have mitigated the effects of political uncertainty on prices. Depression worldwide could have higher explanatory power on the price fluctuations. Furthermore, investors and consumers could have been barely aware of the outcomes of the occupation and the civil war in Anatolia, leading to a lower sensitivity of markets (Hanedar et al. 2016, 2018, 2019). Moreover, we observe that price fluctuations and uncertainty could not end promptly after the establishment of the new regime and the end of the military conflicts, which could mean that the political instabilities have long-lasting effects on the economy and prices. Our findings on the link between higher risk and price fluctuations are consistent with the limited literature on the positive relationship between political

risk and commodity prices (Asteriou and Price 2001; Hou et al. 2020). For instance, Hou et al. (2020) pointed out that political uncertainty is related to lower commodity prices through decreasing demand of consumers and firms for goods and inputs. In our case, we find increasing prices of goods in case of political uncertainty. This increase could be explained by a supply shortage and lower trade, while political uncertainties could have discouraged trade and production. As a final point, besides adding new empirical evidence, our paper adds to the literature (Bittlingmayer 1998; Wang 2019; Wang and Boatwright 2019) using historical episodes as natural experiments to test the impacts of political uncertainty. Our study has some concerns on bias in the estimation of the effects of risks because of other covariates which we could not put our models due to data constraints. Then we exploit a natural experiment experienced in the Ottoman Empire and the Republic of Turkey between 1918 and 1923 to set exogenous political shocks, providing an unbiased causal link from political risks to good prices.

## 5 Conclusion

In this chapter, we attempt to understand the types of political problems that are risky in terms of instability in commodity trading and prices. The causal link is often suffering from bias due to the other covariates of political risk and prices which are not included in our model due to data constraints. To address this issue, using unique price data of İstanbul during the end of the Great War, we present a counterfactual analysis and natural experiment. This methodology is based on the differences in goods' prices between days of high and low political uncertainty. To obtain a convincing causal estimate, this methodology is recently used in the literature about the effects of political risks on financial markets, showing that political problems would lead to volatility through disrupting trade. On the other hand, the literature does not provide detailed information on the relationship between political uncertainty and commodity markets. Our estimation fills the gap in the literature that aims to detect political risk due to omitted factors, by using a natural experiment from the early twentieth century.

We find that only one type of political uncertainty, namely the government resignations, might be related to serious fluctuations in good prices. The increasing volatility of prices could reflect that higher uncertainty on the future of the economy due to weak institutions in case of government turmoil might lead to lower production and trade through the willingness of traders and investors. Besides, our results introduce new evidence for the high response of energy market outcomes to this political problem much during the same period. This effect could be interpreted as evidence for the positive relationship between the lower survival probability of a government and the security of import links, as energy products were mainly obtained abroad. Our findings do not suggest that the political events between 1919 and 1923 created an unstable price system. So, we can conclude that in other political instabilities, such as the dissolution of a country and civil conflicts, political uncertainties do not

necessarily pose more unstable prices. Moreover, we find that the birth of a new regime or institution could not solve problems of food and energy supply promptly. These results could also mean that a good market could not evaluate all events as keys creating price changes.

To sum up, the econometric analysis based on the historical episode of our study reveals that political events played an important role in commodity price fluctuations. The Ottoman case is important because the country experienced an exogenous political shock, alleviating the bias because of omitted variables. However, the insufficient number of observations and the absence of many control variables in the counterfactual model would lead to future research by obtaining more data. Under the lack of data, the policy suggestion of our paper is still crucial for dealing with the relationship between political instability and good price fluctuations, which could herald social tensions over time. In particular, we could say that the uncertainties caused by lower government durability might bring crucial damages to the market mechanism and social order during serious political crises.

**Acknowledgements** I thank Elmas Yıldız Hanedar, Tuğ İnce, and Sezgin Uysal for their helpful comments and materials.

## References


- Anderson JE, Marcouiller D (2002) Insecurity and the pattern of trade: an empirical investigation. *Rev Econ Stat* 84(2):342–352. <https://doi.org/10.1162/003465302317411587>
- Asteriou D, Price S (2001) Political instability and economic growth: UK time series evidence. *Scott J Polit Econ* 48(4):383–399. <https://doi.org/10.1111/1467-9485.00205>
- Baker SR, Bloom N, Davis SJ (2016) Measuring economic policy uncertainty. *Q J Econ* 131(4):1593–1636
- Bittingmayer G (1998) Output, stock volatility, and political uncertainty in a natural experiment: Germany, 1880–1940. *J Financ* 53(6):2243–2257. <https://doi.org/10.1111/0022-1082.00090>
- Bouoiyour J, Selmi R, Hammoudeh S, Wohar ME (2019) What are the categories of geopolitical risks that could drive oil prices higher? Acts or threats?. *Energy Econ* 84:104523
- Çavdar T (1983) *Yüz yıllık pahalılık. Ülke yayınları*, Ankara
- Ediger VŞ (2007) *Osmanlı'da nefit ve petrol. ODTÜ yayıncılık*, Ankara
- Eldem V (1994) Harp ve mütareke yıllarında Osmanlı İmparatorluğu'nun ekonomisi. *Türk Tarih Kurumu basımevi*, Ankara
- Estevadeordal A, Frantz B, Taylor AM (2003) The rise and fall of world trade, 1870–1939. *Q J Econ* 118(2):359–407. <https://doi.org/10.1162/003355303321675419>
- Fromkin D (2001) *A peace to end all peace: the fall of the Ottoman Empire and the creation of the modern Middle East*. Macmillan, New York
- Hanedar AÖ (2018) Bir iktisadi tetikleyici olarak birinci dünya savaşı'nın bitişi: barışın osmanlı imparatorluğu'ndaki etkilerinin iktisadi incelenmesi. *İnsan ve toplum* 8(4):31–56. <https://doi.org/10.12658/M0248>
- Hanedar AÖ (2016) Effects of wars and boycotts on international trade: evidence from the late Ottoman Empire. *Int Trade J* 30(1):59–79. <https://doi.org/10.1080/08853908.2015.1102107>
- Hanedar AÖ, Gencer HG, Demiralay S, Altay İ (2019) The Ottoman dissolution and the İstanbul bourse between war and peace: a foreign exchange market perspective on the Great War. *Scand Econ Hist Rev* 67(2):154–170. <https://doi.org/10.1080/03585522.2018.1546615>

- Hanedar AÖ, Hanedar EY, Torun E (2016) The end of the Ottoman Empire as reflected in the İstanbul bourse. *Hist Methods J Quant Interdiscip Hist* 49(3):145–156. <https://doi.org/10.1080/01615440.2015.1118365>
- Hanedar AÖ, Hanedar EY, Torun E, Ertuğrul HM (2018) Dissolution of an empire: insights from the İstanbul Bourse and the Ottoman War Bond. *Def Peace Econ* 29(5):557–575. <https://doi.org/10.1080/10242694.2016.1239319>
- Hornbeck R, Naidu S (2014) When the levee breaks: black migration and economic development in the American South. *Am Econ Rev* 104(3):963–990
- Hou K, Tang K, Zhang B (2020) Political uncertainty and commodity markets. *Fish Coll Bus Work Pap* (2017-03):025. <https://doi.org/10.2139/ssrn.3064295>
- Jens CE (2017) Political uncertainty and investment: causal evidence from US gubernatorial elections. *J Financ Econom* 124(3):563–579
- Julio B, Yook Y (2012) Political uncertainty and corporate investment cycles. *J Finance* 67(1):45–83
- Kalemli-Özcan S, Nikolsko-Rzhevskyy A, Kwak JH (2020) Does trade cause capital to flow? Evidence from historical rainfall. *J Dev Econ* 147:102537. <https://doi.org/10.1016/j.jdeveco.2020.102537>
- Redding SJ, Sturm DM (2008) The costs of remoteness: evidence from German division and reunification. *Am Econ Rev* 98(5):1766–1797
- Shen Y (2020) Measuring macroeconomic uncertainty: a historical perspective. *Econ Lett* 196:109592. <https://doi.org/10.1016/j.econlet.2020.109592>
- Su CW, Qin M, Tao R, Moldovan NC (2019) Is oil political? From the perspective of geopolitical risk. *Def Peace Econom* 1–17
- Shiller RJ (2017) Narrative Economics. *Ame Econ Rev* 107(4):967–1004. <https://doi.org/10.1257/aer.107.4.967>
- Tasvir-i Efkâr (1918, December 16) Gala-yi es'ar, p 1
- Toprak Z (1982) Türkiye'de "milli iktisat" (1908–1918). *Yurt yayınları*, Ankara
- Wang H (2019) The causality link between political risk and stock prices: a counterfactual study in an emerging market. *J Financ Econ Policy* 11(3):338–367. <https://doi.org/10.1108/JFEP-07-2018-0106>
- Wang H, Boatwright AL (2019) Political uncertainty and financial market reactions: a new test. *Int Econ* 160:14–30. <https://doi.org/10.1016/j.inteco.2019.07.004>
- Wang T, Zhang D, Broadstock DC (2019) Financialization, fundamentals, and the time-varying determinants of US natural gas prices. *Energy Econ* 80:707–719. <https://doi.org/10.1016/j.eneco.2019.01.026>
- Yin L, Han L (2014) Macroeconomic uncertainty: does it matter for commodity prices? *Appl Econ Lett* 21(10):711–716. <https://doi.org/10.1080/13504851.2014.887181>
- Zhu H, Huang R, Wang N, Hau L (2020) Does economic policy uncertainty matter for commodity market in China? Evidence from quantile regression. *Appl Econ* 52(21):2292–2308. <https://doi.org/10.1080/00036846.2019.1688243>



# A Nonlinear Panel ARDL Analysis of Pollution Haven/Halo Hypothesis



Ebru Çağlayan-Akay  and Zamira Oskonbaeva 

**Abstract** There is a growing popularity for the nonlinear econometric approaches, since linkages among variables are not always linear. Nonlinear approaches provide a broader range of knowledge compared to the linear model. This research aims to assess the impact of foreign direct investment on pollution. To capture the potential asymmetries resulting from rise and fall in the foreign direct investments, the nonlinear panel autoregressive distributed lag approach is employed. In the empirical analysis, annual data of selected 22 transition economies from 1995 to 2016 is utilized. The findings highlighted the existence of asymmetric linkages among variables. In other words, evidence reveals that positive shock in foreign direct investment improves environmental quality, while the negative shock is detrimental to the environment.

**Keywords** FDI · Pollution halo hypothesis · CO<sub>2</sub> emissions · Panel nonlinear ARDL · Pollution haven hypothesis

## 1 Introduction

Economic theories are mostly run based on linearity assumptions about the association among the underlying variables or the common normality assumptions of their distributions (Kotchoni 2018). However, it is more realistic and less restrictive to conclude that the linkages among several economic variables are best defined by a nonlinear framework. Many nonlinear models can give better economic assessment and have widely been implemented by applied economists (Kim et al. 2004).

If linear models provide an acceptable interpretation of most linkages among economic variables, then standard procedures are reasonable, and the stylized facts

---

E. Çağlayan-Akay  
Department of Econometrics, Marmara University, Istanbul, Turkey  
e-mail: [ecaglayan@marmara.edu.tr](mailto:ecaglayan@marmara.edu.tr)

Z. Oskonbaeva (✉)  
Department of Economics, Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan  
e-mail: [zamira.oskonbaeva@manas.edu.kg](mailto:zamira.oskonbaeva@manas.edu.kg)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_11](https://doi.org/10.1007/978-3-030-85254-2_11)

they generate give a strong foundation for policymaking and theorizing. However, if there are major nonlinearities, these should be integrated into the macro models and the stylized facts produced by the linear models should be brought into question (Koop and Potter 2001). To put it another way, when dealing with nonlinear data generation systems, linear procedures can lead to false conclusions (Yavuz and Yilanci 2012). Adoption of a linear procedure is appropriate if the linearity test fails to find proof of nonlinearity (Liew et al. 2003). Since preferences are considered to be convex and the economy is vulnerable to different exogenous random shocks, economic systems are essentially considered to be stochastically nonlinear (Hinich et al. 2005). Macroeconomic indicators did not behave in the same manner during the ups and downs. The rising trend has abruptly and rapidly changed to a weakening trend. Consequently, the rises and decreases occurred in the explanatory variable can influence the dependent variable in a different manner (Keynes 1936). Recent studies suggest that financial or economic variables can suffer from inherent nonlinearities (Atil et al. 2014; Shin et al. 2014). Besides, existing theories and findings of the researches indicate that major economic activities exhibit nonlinear pattern. For instance, several theoretical macroeconomic models are highly nonlinear (Grandmont 1985; Hall 1990). Furthermore, extensive empirical research has indicated statistical evidence of considerable nonlinearity in the generation of essential macroeconomic indicators (Altug et al. 1999, among others).

The ongoing debate on asymmetric and nonlinear models starts with the contribution of Balke and Fomby (1997) who developed the threshold co-integration with a regime-switching type model. The procedure advanced by Granger and Yoon (2002) is also well documented in the literature. Their paper was advanced by Schorderet (2003) who evaluated the asymmetrical impact of hidden co-integration. Recently, Shin et al. (2014) developed a nonlinear model inspired by the study of Pesaran et al. (2001).

The assumption that accurate estimation of parameters and reliable statistical inference depends mainly on the suitable procedure is hardly controversial. It is therefore essential to confirm whether or not a nexus between variables appears to be generated by a linear model against an alternative that is not linearly associated. This study considers nonlinear panel ARDL procedure to explore the potential asymmetries resulting from changes occurred in FDI on the environment. Nonlinearity and co-integration among the fundamental variables can be assessed concurrently. Moreover, this procedure has advantages over other co-integration techniques. First, it gives reliable outcomes in the case of small sample sizes. Second, this specification is a flexible one that can be employed with variables with different orders of integration.

Foreign direct investment (FDI) is essential to many countries' economic development. Moreover, human capital and financial sector development are impacted by FDI inflows. It is well documented that in addition to economic benefits, there are also environmental impacts of FDI (Jensen 1996, among others). FDI-environment relationship has attracted a significant amount of attention from researchers and policy-makers. There exist two key points of view regarding the FDI-pollution nexus: pollution haven and pollution halo hypothesis. First view postulates that FDI flows

are detrimental to the environment, since it is responsible for shifting dirty industries to developing countries. As a result of their low labor costs and natural resources, as well as their weak environmental regulations, developing countries are seen as a pollution haven for advanced economies. The second hypothesis states that FDI is beneficial for host countries since it brings clean and advanced technology which in turn leads to a drop in carbon emissions. Also, it implies that foreign investors conduct business in an environmentally friendly manner.

Researchers attempted to validate the above-mentioned hypotheses at a single country or regional level. Mixed outcomes were achieved. The primary goal of current research is to assess the validity of the two hypotheses in the context of transition economies, considering the aforementioned concerns. For this purpose, annual data of selected transition economies for the period 1995–2016 were utilized. Since FDI is more vulnerable to shocks its impact on the environment might be different too. Changes occurred in FDI may impact environment in different manner. So, to check long-term and short-term asymmetries in the FDI-pollution nexus panel nonlinear ARDL model was applied.

The reason why we decided to explore the FDI-pollution nexus in transition economies is that firstly, the fall of the Socialist regime created countless prospects for investment in the countries under consideration. FDI was also considered as a catalyst since it could bring the advanced techniques and new innovations to host countries. During the transition process scarcity of domestic resources made the above-mentioned economies dependent on the attraction of foreign direct investment. Secondly, Russian Federation, Poland, and Kazakhstan were among the largest emitters in 2018 (Union of Concerned Scientists Data 2020). Therefore, the issue of FDI-pollution in the context of transition economies requires a thorough examination.

The paper's key contribution to the earlier studies can be seen in three aspects. First, it elaborates the validity of aforementioned hypotheses in the context of transition economies by utilizing data for longer periods and more countries. Second, the analysis in this study is focused on the nonlinear nature of the FDI-pollution nexus. It shows whether positive shocks in the FDI value influence carbon emissions differently from negative ones. The nonlinearity of the FDI-pollution nexus has important policy consequences and sheds light on the added features of environmental impact of FDI in the economies described above. Third, the panel nonlinear ARDL approach was employed to explore the existence of an asymmetric association among variables. No study before has employed this approach to capture the impact of changes occurred in FDI flows on pollution in the selected transition economies.

## 2 Literature

A great deal of researches have been conducted to determine the reasons for environmental damage (Işik et al. 2017). Particular attention has been paid to FDI-pollution relationship among researchers. Empirical evidence indicates that there exist two contradictory views regarding the FDI-pollution nexus. The results of some studies

confirm the FDI-pollution haven hypothesis, while the findings of other studies indicate that pollution halo hypothesis is supported (Rafindadi et al. 2018; Öztürk and Öz 2016; Balsalobre-Lorente et al. 2019, among others). Outcomes of these researches are presented in Table 1.

As can be observed from Table 1 less attention has been paid to the experience of transition economies. Moreover, a common feature of prior researches is that they have presumed the effect of foreign direct investment on the pollution to be symmetric (except Rahman et al. 2019). Unlike prior studies, we utilize the nonlinear panel ARDL procedure to assess the asymmetric relations among aforementioned variables. This study, therefore, extends prior studies to identify the asymmetric relations between the variables in the selected transition economies.

### 3 Variables and Data Set

In this research paper annual data of 22 selected transition economies namely, Albania, Armenia, Belarus, Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Kazakhstan, Kyrgyz Republic, Latvia, Lithuania, Moldova, Poland, Romania, Russian Federation, Slovak Republic, Slovenia, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan over the period of 1995–2016 are used. Our choice of countries and period are dependent on data availability. Utilized data is collected from the World Bank and U.S. Energy Information databases. Descriptions of the data are displayed in Table 2.

As can be seen from Table 2 variables are used in logarithms. The logarithm is denoted as L. The main model used is described as follows:

$$LCO_{2it} = \alpha_0 + \alpha_1 LFDI_{it} + \alpha_2 LPOP_{it} + \alpha_3 LEC_{it} + \alpha_4 LURB_{it} + \varepsilon_{it} \quad (1)$$

where  $LCO_2$  denotes the dependent variable which is used as an indicator of pollution.  $LFDI_{it}$ ,  $LPOP_{it}$ ,  $LEC_{it}$ , and  $LURB_{it}$  are explanatory variables.  $\varepsilon_{it}$  represents the error term.  $t$  and  $i$  indices indicate time and country, respectively. Population, energy consumption, and urbanization are included as additional repressors.

### 4 Methodology: Nonlinear Panel ARDL Model

Since our study has focused attention on asymmetric effect of foreign direct investment flows on pollution, in this section are discussed procedures utilized in the empirical analysis. This section will describe steps to be taken to investigate the nexus between FDI-pollution. Firstly, we consider the developing stages and main assumptions of nonlinear panel ARDL framework. Then preliminary analysis conducted is presented.

**Table.1** Summary and major findings of FDI-pollution nexus

Author	Country	Method	Outcomes
Nguyen et al. (2020)	33 economies 1996–2014	STIRPAT model	Pollution haven
Guzel and Okumus (2020)	ASEAN-5 countries	CCEMG and AMG estimators	Pollution haven
Mert and Caglar (2020)	Turkey	Hatemi J-Irandost hidden co-integration	Pollution halo
Rahman et al. (2019)	Pakistan 1975–2016	NARDL technique	Pollution haven
Balsalobre-Lorente et al. (2019)	MINT countries 1990–2013	Pedroni, Kao, and Fisher for co-integration; FMOLS, DOLS, and Dumitrescu-Hurlin causality	Pollution halo
Shahbaz et al. (2018)	France 1955–2016	Bootstrap ARDL	Pollution haven
Rafindadi et al. (2018)	GCC countries 1990–2014	PMG methodology	Pollution halo
Solarin et al. (2017)	Ghana 1980–2012	ARDL	Pollution haven
Sun et al. (2017)	China 1980–2012	ARDL	Pollution haven
Bakirtaş and Çetin (2017)	MIKTA countries 1982–2011	Panel VAR	Pollution haven
Sapkota and Bastola (2017)	14 Latin American countries	Time series analysis	Pollution haven
Öztürk and Öz (2016)	Turkey 1974–2011	Maki for co-integration, DOLS, and Granger causality	Pollution halo
Mert and Bölük (2016)	Kyoto countries	Unbalanced panel ARDL	Pollution halo
Baek (2016)	ASEAN 5 countries 1981–2010	Pedroni for co-integration and panel ARDL	Pollution haven
Hao and Liu (2015)	29 provinces of China 1995–2011	FE and GMM approaches	Pollution halo
Tang and Tan (2015)	Vietnam 1976–2009	Johansen co-integration and Granger causality	Pollution halo
Lau et al. (2014)	Malaysia 1970–2008	ARDL and Granger causality	Pollution haven
Merican et al. (2007)	ASEAN 5 countries 1970–2001	ARDL	Mixed results

ASEAN (The Association of Southeast Asian Nations), MINT (Mexico, Indonesia, Nigeria, and Turkey), GCC (The Gulf Cooperation Council) MIKTA (Mexico, Indonesia, the Republic of Korea, Turkey, and Australia), STIRPAT (Stochastic (ST) estimation of environmental impacts (I) by regression (R) on population (P), affluence (A) and technology (T)), CCEMG (Common Correlated Effects Mean Group), AMG (Augmented Mean Group), NARDL (Nonlinear Autoregressive Distributed Lag), DOLS (Dynamic Ordinary Least Squares), FE (Fixed Effect), and GMM (Generalized Method of Moments)

**Table.2** Variables

Variables	Abbreviation	Description
Carbon dioxide emission	LCO <sub>2</sub>	CO <sub>2</sub> emissions (kg per 2010 US\$ of gross domestic product)
Foreign direct investment	LFDI	FDI as a percent of gross domestic product
Population	LPOP	Population (total)
Energy consumption	LEC	EC (in quadrillion British thermal units)
Urbanization	LURB	Urban population (the population living in cities)

Panel ARDL model which can be described as a dynamic panel heterogeneity analysis was developed by Pesaran et al. (1999, 2004). Panel ARDL (p,q) procedure can be expressed as shown in Eq. (2):

$$y_{it} = \sum_{j=1}^p \gamma_{ij} y_{i,t-j} + \sum_{j=0}^q \delta'_{ij} x_{i,t-j} + u_{it} \tag{2}$$

where  $x_{it}$  ( $k \times 1$ ) indicate a vector of explanatory variables,  $\gamma_{ij}$  are scalars which show the parameters of the lagged explained variable,  $\delta_{ij}$  are  $k \times 1$  coefficient vectors.  $p$  denotes the lags of the dependent variable and  $q$  represents the lags of the explanatory variables.

In our study we can estimate the following model as described by Eq. (3):

$$LCO_{2it} = \alpha_i + \sum_{l=1}^p \beta_0 LCO_{2i,t-l} + \sum_{l=0}^q \beta_1 LFDI_{i,t-l} + \sum_{l=0}^q \beta_2 X_{i,t-l} + u_{it} \tag{3}$$

An error correction term can be included. In this case, Eq. (3) can be expressed in the following way:

$$\begin{aligned} \Delta LCO_{2it} = & \alpha_i + \Phi_i(LCO_{2i,t-l} - \theta_1 LFDI_{i,t-l} - \theta_2 X_{i,t-l}) + \sum_{l=1}^{p-1} \lambda_1 \Delta LCO_{2i,t-l} \\ & + \sum_{l=0}^{q-1} \lambda_2 \Delta LFDI_{i,t-l} + \sum_{l=0}^{q-1} \lambda_3 \Delta X_{i,t-l} + u_{it} \end{aligned} \tag{4}$$

where country and time can be indicated by  $i$  and  $t$ , respectively.  $LCO_2$  is the dependent variable.  $LFDI$  is our main variable of interest and other control variables are shown by  $X$ . Short-term coefficients of the lagged dependent variable,  $FDI$ , and other control variables are indicated by the notation  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , respectively.  $\Theta_1$  and  $\Theta_2$  represent the long-term parameters for  $FDI$  and other control variables, respectively.

$\Phi_i$  shows the adjustment coefficient,  $u_{it}$  denotes residual. Equation (4) can be estimated with the help of widely used estimators such as the mean group estimator of Pesaran and Smith (1995), pooled mean group estimator proposed by Pesaran et al. (1999), and dynamic fixed effects (FE) estimator. The long-term equilibrium and the heterogeneity of the dynamic adjustment process are the main assumptions of the above-mentioned estimators. In order to identify the basic characteristics of the above-described estimators, the assumptions relating to each estimator are presented below.

As can be seen from Eqs. (3) and (4), FDI was not decomposed into positive and negative changes. Since in the panel ARDL framework the influence of FDI on the environment is assumed to be symmetric. The asymmetric impact of FDI is estimated by implementing a nonlinear panel ARDL model. When the reaction to shocks in one stage of the economic cycle varies from the reaction in the other, this is depicted as asymmetric behavior. Shocks in FDI flows may stem from different reasons: political instability, changing macroeconomic conditions in a country, etc. In that case, asymmetric fluctuations cannot be captured by traditional linear models (Canepa et al. 2019; Sichel 1993). Therefore, by utilizing a nonlinear procedure in identifying such asymmetry can be obtained consistent outcomes. A nonlinear panel ARDL procedure was introduced based on the framework of Pesaran et al. (1999) and Pesaran et al. (2001). Although the procedure was developed for time series, for panel data can be formulated as below:

$$\begin{aligned} \Delta y_{it} = & \Phi_i \left( y_{i,t-j} - \theta_1 x_{i,t-j}^+ - \theta_2 x_{i,t-j}^- \right) + \sum_{j=1}^{p-1} \lambda_1 \Delta y_{i,t-j} \\ & + \sum_{j=0}^{q-1} \left( \lambda_2 \Delta x_{i,t-j}^+ + \lambda_3 \Delta x_{i,t-j}^- \right) + u_{it} \end{aligned} \tag{5}$$

The model decomposes independent variable into negative and positive changes (Shin et al. 2014). This procedure can be applied if it is believed that dependent variable responds in different manners to the shocks occurred in independent variable.

In our study the method we utilized allows for the asymmetric response of carbon emissions to FDI. To put it another way, positive and negative FDI shocks are unlikely to have the same effect on carbon emissions. The nonlinear panel ARDL model can be expressed as described by Eq. (6):

$$\begin{aligned} \Delta y_{it} = & \Phi_i \left( y_{i,t-j} - \theta_1 x_{i,t-j}^+ - \theta_2 x_{i,t-j}^- \right) + \sum_{j=1}^{p-1} \lambda_1 \Delta y_{i,t-j} \\ & + \sum_{j=0}^{q-1} \left( \lambda_2 \Delta x_{i,t-j}^+ + \lambda_3 \Delta x_{i,t-j}^- \right) + u_{it} \end{aligned} \tag{6}$$

where  $i$  and  $t$  denote country and time, respectively,  $LCO_2$  is our dependent variable,  $LFDI^+$  and  $LFDI^-$  denote the positive and negative FDI shocks, respectively,  $X$  represents a group of other control variables: energy consumption, urbanization and population. Notation  $\lambda_1, \lambda_2, \lambda_3$ , and  $\lambda_4$  are the short-term coefficients of the lagged dependent variable, FDI and other control variables, respectively.  $\theta_1, \theta_2$  and  $\theta_3$  are long-run coefficients for FDI and other control variables. Lastly,  $\Phi_i$  represents error correction term.  $\Phi_i$  should take statistically significant negative value, lower than one. If the value of  $\Phi_i$  is greater it implies that after a shock occurred in short-term, long-run equilibrium can be reached quickly (Pesaran et al. 2001).  $u_{it}$  represents the error term.

The key benefit of this technique is that it can be applied in case when the integration order of data is mixed. It is well documented that other approaches require the integration order of the series to be similar. Moreover, this procedure can be applicable for seizing nonlinearities that occur due to shocks in macroeconomic variables (Mihajlović and Marjanović 2020; Mensi et al. 2017). This approach tackles the multi-collinearity problems regarding the data by selecting the most appropriate lag order for the variables under consideration (Shin et al. 2014).

Shocks can be measured as positive and negative partial sum decompositions of FDI changes and can be expressed as follows:

$$LFDI_{i,t}^- = \sum_{j=1}^t \Delta LFDI_{ij}^- = \sum_{j=1}^t \min(\Delta LFDI_{ij}, 0) \tag{7}$$

$$LFDI_{i,t}^+ = \sum_{j=1}^t \Delta LFDI_{ij}^+ = \sum_{j=1}^t \max(\Delta LFDI_{ij}, 0) \tag{8}$$

where  $\max$  and  $\min$  represent positive and negative changes in FDI. If coefficients  $LFDI^+$  and  $LFDI^-$  are substantially different from each other, there is evidence of asymmetry, otherwise their effect on pollutions is considered to be the same.

The model we applied in this study can be estimated with the help of the Mean Group (MG) and Pooled Mean Group (PMG) estimators.

Following Pesaran et al. (1999) PMG estimator can be computed as described by Eqs. (9) and (11):

$$\hat{\theta} = - \left\{ \sum_{i=1}^N \frac{\hat{\phi}_i^2}{\hat{\sigma}_i^2} X_i' H_i X_i \right\}^{-1} \times \left\{ \sum_{i=1}^N \frac{\hat{\phi}_i}{\hat{\sigma}_i^2} X_i' H_i (\Delta y_i - \hat{\phi}_i y_{i,-1}) \right\} \tag{9}$$

$$\hat{\phi}_i = (\hat{\xi}_i' H_i \hat{\xi}_i)^{-1} \hat{\xi}_i' H_i \Delta y_i \tag{10}$$

$$\hat{\sigma}_i^2 = T^{-1} e (\Delta y_i - \hat{\phi}_i \hat{\xi}_i)' H_i (\Delta y_i - \hat{\phi}_i \hat{\xi}_i) \tag{11}$$



where  $\hat{\xi}_i = y_{i,-1} - X_i \hat{\Theta}$ .

MG estimator can be computed as below:

$$\hat{\phi}_{MG} = N^{-1} \sum_{i=1}^N \hat{\phi}_i, \tag{12}$$

with the variance

$$\hat{\Delta}_{\hat{\phi}} = \frac{1}{N(N-1)} \sum_{i=1}^N (\hat{\phi}_i - \hat{\phi})^2 \tag{13}$$

Firstly, we adopted both MG and PMG estimators. Parameter heterogeneity is the main assumption of the MG estimator. In other words, all parameters can be different across countries. Long-term estimates are obtained by averaging the individual country-specific coefficients. Having a large enough N and T, however, is a necessary condition for this estimator’s accuracy and reliability. This estimator can be susceptible to outliers if the number of observations is small (Favara 2003). Furthermore, the estimator described above does not account for cross-sectional dependence.

The pooled mean group estimator’s prominent characteristic is that it requires long-term coefficients to be homogeneous across countries while permitting short-term parameters and error variances to differ. The above-mentioned estimator requires the following assumptions to behold: (i) error terms are not serially correlated; (ii) a long-term relationship exists between the variables; (iii) and long-term parameters should be identical across groups. Violation of these assumptions will lead to inconsistent estimation in PMG. Since the consequences of financial collapse and external shocks differ by region, short-term adjustments are anticipated to be country-specific. Furthermore, this estimator can be depicted as less susceptible to the presence of outliers in small sample sizes (Pesaran et al. 1999).

To choose an efficient estimator has been adopted the Hausman test. Acceptance of the null hypothesis postulates that the PMG estimator is efficient, while the MG estimator is considered to be robust and efficient in case of rejection of the null hypothesis (Blackburne and Frank 2007). Symmetries can be tested by utilizing Wald test (Shin et al. 2014). The null hypothesis  $\theta_1 = \theta_2$  can be used to test long-run symmetry. The null hypothesis  $\lambda_2 = \lambda_3$  can be used to determine whether or not there is short-term symmetry. If the Wald tests’ F-statistics are non-significant, it implies that asymmetry behavior does not exist among series.

There are many similarities between the dynamic FE and pooled mean group estimator. In the long run, however, the co-integrating vector coefficient is assumed to be the same across all groups. Moreover, the speed of convergence and the short-term coefficients are also restricted to be identical. Since the error term and the lagged dependent variable are endogenous, biases can occur. In other words, simultaneous equation bias is the main drawback of the FE model (Baltagi et al. 2000).

## 4.1 Preliminary Analysis

Before conducting the main estimations cross-sectional dependence (CD) and stationary tests are performed. In the first stage issues of cross-sectional dependence should be addressed. Disturbances are believed to be cross-sectionally independent in panel data analysis. However, due to geographical proximity, political or economic drivers may arise the concerns relating to cross-sectional linkage (Gaibullov et al. 2014). Cross-sectional dependence issues may emerge as a result of spatial or spillover impacts, or unobservable common factors (Baltagi and Pesaran 2007). Analysis of the macroeconomic variables for various countries is influenced by common significant events that could contribute to dependency among panel units. In several ways, the transition economies are interlinked. Worldwide fluctuations, such as oil price instability and economic turmoil, as well as specific domestic or sectoral shocks, are likely to affect these economies concurrently. The existence of common shocks can cause interdependence among the groups in the panel (Munir and Kok 2015). Stationary tests can lead to biased inferences if correlation between units is not taken into consideration (Pesaran 2007). Hence, it is essential to assess whether or not there is correlation between units.

The empirical analysis begins with the CD test which was introduced by Pesaran (2004). The outcomes of the CD test allow us to utilize either first-generation or second-generation panel unit root tests. The former does not account for cross-sectional dependencies between entities, while the latter does. There are two main points to consider in terms of cross-sectional dependency's presence or absence. First, if there exists cross-correlation, traditional stationary tests, which are based on the presumption of cross-sectional independence, are prone to large-scale distortions (O'Connell 1998; Maddala and Wu 1999). Second, the lack of cross-sectional dependence will cause significant power losses in stationary tests that allow cross-correlation. Therefore, it's necessary to select appropriate stationary tests.

The major feature of this procedure is its applicability in both cases when  $T > N$  and  $N > T$ . Test statistics can be calculated as follows:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \hat{\rho}_{ij} \quad (14)$$

$\hat{\rho}_{ij}$  denotes the simple correlation coefficient between the residuals derived from the estimation of each equation using the least squares method.

After examining cross-sectional dependence between series, in order to ensure that the series do not contain the unit root, panel unit root procedure is employed. We utilized second-generation and first-generation panel unit root tests proposed by Pesaran (2007) and Levin et al. (2002).

Pesaran (2007) Cross-sectional Augmented Dickey Fuller (CADF) regression equation can be described as below:

$$\Delta Y_{it} = \alpha_i + b_i Y_{i,t-1} c_i \bar{Y}_{t-1} + d_i \Delta \bar{Y}_t + \varepsilon_{it} \quad (15)$$

where  $\bar{Y}_t$ , refers to the average of all cross-sectional observations over time. In the case of autocorrelation, the above Eq. (15) can be extended as follows:

$$\Delta Y_{it} = \alpha_i + \rho_i^* Y_{it-1} + d_0 \bar{Y}_{t-1} + \sum_{j=0}^p d_{j+1} \Delta \bar{Y}_{t-j} + \sum_{k=1}^p c_k \Delta Y_{i,t-k} + \varepsilon_{it} \quad (16)$$

After estimation of CADF regression CIPS (Cross-sectional augmented Im Pesaran Shin) statistic can be obtained as shown in Eq. (17):

$$CIPS = \frac{1}{N} \sum_{i=1}^N CADF_i \quad (17)$$

where  $CADF_i$  represents the means of the t-statistics of the lagged variables.

Levin et al. (LLC 2002) test can be expressed as follows:

$$\Delta y_{it} = \alpha_0 + \rho y_{it-1} + \varepsilon_{it} \quad (18)$$

where  $\rho$  is assumed to be identical and fixed across units.  $\varepsilon_{it}$  denotes the error process that is spread through variables independently. LLC is a procedure which analyze the impact of deterministic trend. If  $\rho$  is equal to zero it implies that the series are not stationary. In this case in order to make series stationary the difference must be obtained. Under the null hypothesis that  $\delta = 0$ , LLC (2002) t-test statistic is calculated as below:

$$t_\delta = \frac{\hat{\delta}}{std(\hat{\delta})} \quad (19)$$

## 5 Empirical Findings

In this section are presented empirical findings of our study. The outcomes of CD test are displayed in Table 3.

**Table.3** Cross-sectional dependence test

Variables	LCO <sub>2</sub>	LFDI	LEC	LPOP	LURB	LURB <sup>2</sup>
Test statistics	53.04	16.37	6.86	3.17	1.70	1.74
Probability	0.000	0.000	0.000	0.002	0.089	0.081

Null hypothesis is  $H_0: \rho_{ij} = c(u_{it}, u_{jt}) = 0 \ i \neq j$  and it implies the lack of cross-correlation among units

**Table.4** Panel unit root tests

		tbar statistic	zbar statistic	Probability	tbar statistic	zbar statistic	Probability
		Level			First difference		
Pesaran	LCO <sub>2</sub>	-2.188	0.615	0.731	-3.237*	-4.673*	0.000
	LFDI	-2.664**	-1.784**	0.037			
	LEC	-1.974	1.694	0.955	-3.689*	-6.954*	0.000
	LPOP	-2.724**	-2.089**	0.018			
		Statistic					
LLC	LURB	-8.1602*		0.000			
	LURB <sup>2</sup>	-8.1902*		0.000			

Null hypothesis indicates that series are not stationary. \* and \*\* denote significance at the 1% and 5% levels, respectively

The findings illustrate that cross-correlation exists between all the series (except urbanization rate and urbanization (quadratic term)). Given the above-described results, the stationary analysis is performed. The obtained findings are summarized in Table 4.

The results indicate that series exhibit stationarity at levels (except energy consumption and carbon emissions). In other words, the order of integration is found to be mixed I(0) and I(1). These results demonstrate that we can move forward with the panel nonlinear ARDL approach. Since this approach enables us to utilize both I(0) and I(1).

According to the Hausman test (1978) results the PMG estimator was found to be an efficient estimator for modeling FDI-pollution nexus. Thus, only the outcomes obtained from the PMG estimator are discussed in this paper. The results are presented in Table 5.

Findings indicate that coefficient of foreign direct investment increases is negative and FDI decreases is positive. To put it differently, positive shocks that occur in FDI adversely impact the environment. Carbon emissions will be decreased by 0.3688 percent for every 1% rise in FDI. Negative shocks occur in FDI on the contrary, escalate carbon emissions. Carbon dioxide emissions are expected to rise by 0.3328 percent for every 1% fall in FDI. The magnitude of FDI<sup>+</sup> is found greater in volume than that of FDI<sup>-</sup>. Our empirical results imply that FDI can be beneficial to host countries. Thus, empirical results illustrate that the asymmetric pollution halo hypothesis is confirmed in economies under consideration. These findings suggest that not taking into account the nonlinearity and asymmetry in modeling the nexus between FDI-pollution may lead to misleading conclusions.

The positive and significant coefficient of LEC specifies that energy consumption upsurges CO<sub>2</sub> emissions. It is seen that a 1% escalation in energy consumption is responsible for a 0.15% upsurge in pollutions. The population is also observed to positively and significantly impact environmental degradation. The population coefficient indicates that a 1% rise in population will cause an upsurge in carbon

**Table.5** Results of panel nonlinear ARDL model

Variables	Coefficients	Standard error
Long-term coefficients		
LFDI <sup>+</sup>	-0.3688*	0.0480
LFDI <sup>-</sup>	0.3328*	0.0496
LPOP	11.6570*	1.0538
LEC	0.1462*	0.0468
LURB	10.7636**	5.9150
LURB <sup>2</sup>	-0.6572*	0.2101
Short-term coefficients		
Error correction term	-0.3604*	0.1030
$\Delta$ LFDI <sup>+</sup>	0.0595	0.0610
$\Delta$ LFDI <sup>-</sup>	-0.1625**	0.0833
$\Delta$ LPOP	31.0914**	16.0739
$\Delta$ LEC	0.3126*	0.0828
$\Delta$ LURB	1194.941	2302.912
$\Delta$ LURB <sup>2</sup>	-38.86062	75.8768
Constant	-70.2020*	19.9968
Hausman test	1.65	
P-value	0.9491	
W <sub>LR</sub>	455.56 (0.000)	
W <sub>SR</sub>	3.51 (0.061)	
Log likelihood	843.8456	
Number of observations	462	

The Wald statistics for short- and long-term symmetry under null hypotheses are shown by W<sub>SR</sub> and W<sub>LR</sub>. Probabilities are given in parenthesis. \* and \*\* denote significance at the 1% and 10% levels, respectively

pollution by 11.65%. Turning to LURB and LURB<sup>2</sup>, in Table 5 it is observed that the former is positive while the latter is negative. This outcome reveals the presence of an inverted U-shaped linkage between urbanization and carbon emissions. It indicates that as urbanization expands, carbon emissions initially rise and then start to decrease. It is noteworthy that the magnitude of escalation in emissions associated with urbanization is higher than the potential reduction.

Negative and statistically significant error correction coefficient shows how quickly variables return to equilibrium. It implies that long-term estimates are not spurious.

## 6 Conclusion

This research paper aims to assess the validity of aforementioned two hypotheses in the case of selected transition economies. For this purpose, the panel nonlinear ARDL framework has been employed to the annual data over the period 1995–2016. We separate positive shocks from negative shocks through the partial sum concept and utilize the nonlinear panel ARDL procedure to determine whether FDI inflows have symmetric or asymmetric impacts.

The linkage between foreign direct investment and pollution is found to be asymmetric and adverse. It implies that in the case of transition economies, the pollution halo hypothesis is confirmed. This result is similar to the related literature mentioned above. Effect of foreign direct investment should be taken into consideration while implementing CO<sub>2</sub> mitigation policies. According to the results obtained it can be inferred that FDI can be beneficial to host countries. The negative and significant coefficient of FDI+ specifies that positive shocks can help mitigate CO<sub>2</sub> emissions. Conversely, negative shocks in FDI will increase environmental harm. These findings are of great significance as they suggest that any policy designed to stimulate FDI will result in reduced deterioration of the ecosystem while decreasing FDI will worsen environmental quality. As a result, these findings indicate that analyzing the nexus between FDI and carbon emissions without taking into account nonlinearity and asymmetry can lead to misleading inferences. In light of these findings, a lot of effort should be put into making the country more FDI friendly. The government should prioritize not only the FDI volume but also the efficiency of the FDI. FDI brings green technologies and this creates spillovers for local firms. Local companies can be promoted to learn and adopt advanced technology of international companies and thus, benefit fully from positive spillover effects. Besides, tax incentives can be used as a tool for an increase in FDI flows. Obtained findings will provide new insights into recent trends, thereby giving policy-makers and government officials a better understanding of the nexus between FDI-pollution.

## References

- Altug S, Ashley R, Patterson DM (1999) Are technology shocks nonlinear? *Macroecon Dyn* 3(4):506–533
- Atil A, Lahiani A, Nguyen DK (2014) Asymmetric and nonlinear pass-through of crude oil prices to gasoline and natural gas prices. *Energy Policy Elsevier* 65(C):567–573
- Baek J (2016) A new look at the FDI–income–energy–environment nexus: dynamic panel data analysis of ASEAN. *Energy Policy* 91:22–27
- Bakirtaş I, Çetin MA (2017) Revisiting the environmental Kuznets curve and pollution haven hypotheses: MIKTA sample. *Environ Sci Pollut Res* 24(22):18273–18283. <https://doi.org/10.1007/s11356-017-9462-y>. Epub PMID: 28639013
- Balke N, Fomby T (1997) Threshold cointegration. *Int Econ Rev* 38(3):627–645. <https://doi.org/10.2307/2527284>

- Balsalobre-Lorente D, Gokmenoglu KK, Taspinar N, Cantos-Cantos JM (2019) An approach to the pollution haven and pollution halo hypotheses in MINT countries. *Environ Sci Pollut Res* 26:23010–23026. <https://doi.org/10.1007/s11356-019-05446-x>
- Baltagi BH, Pesaran MH (2007) Heterogeneity and cross section dependence in panel data models: theory and application. *J Appl Economet* 22(2):229–232
- Baltagi BH, Griffin JM, Xiong W (2000) To pool or not to pool: homogeneous versus heterogeneous estimators applied to cigarette demand. *Rev Econ Stat* 82(1):117–126
- Blackburne EF, Frank MW (2007) Estimation of nonstationary heterogeneous panels. *Stand Genomic Sci* 7(2):197–208. <https://doi.org/10.1177/1536867X0700700204>
- Canepa A, Chini EZ, Alqaralleh H (2019) Global cities and local housing market cycles. *J R Estate Financ Econ* 1–27. <https://doi.org/10.1007/s11146-019-09734-8>
- Favara G (2003) An empirical reassessment of the relationship between finance and growth. IMF, Washington, DC
- Gaibulloev K, Sandler T, Sul D (2014) Dynamic panel analysis under cross-sectional dependence. *Polit Anal* 22:258–273. <https://doi.org/10.1093/pan/mpt029>
- Grandmont JM (1985) On endogenous competitive business cycles. *Econometrica* 53:995–1045
- Granger CWJ, Yoon G (2002) Cointegration. University of California, Economics Working Paper No. 2002-02. <https://ssrn.com/abstract=313831>. <https://doi.org/10.2139/ssrn.313831>
- Guzel AE, Okumus I (2020) Revisiting the pollution haven hypothesis in ASEAN-5 countries: new insights from panel data analysis. *Environ Sci Pollut Res* 27:18157–18167. <https://doi.org/10.1007/s11356-020-08317-y>
- Hall R (1990) Invariance properties of Solow's productivity residual. In: Diamond P (ed) *Growth/productivity/employment*. MIT Press, Cambridge
- Hao Y, Liu YM (2015) Has the development of FDI and foreign trade contributed to China's CO<sub>2</sub> emissions? An empirical study with provincial panel data. *Nat Hazards* 76:1079–1091. <https://doi.org/10.1007/s11069-014-1534-4>
- Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46(6):1251–1271
- Hinich MJ, Mendes EM, Stone L (2005) Detecting nonlinearity in time series: surrogate and bootstrap approaches. *Stud Nonlinear Dyn Econo* 9(4):1–13
- Işik C, Kasımatı E, Ongan S (2017) Analyzing the causalities between economic growth, financial development, international trade, tourism expenditure and the CO<sub>2</sub> emissions in Greece. *Energy Sources Part B Econ Plan Policy* 12:665–673. <https://doi.org/10.1080/15567249.2016.1263251>
- Jensen VM (1996) Trade and environment: the pollution haven hypothesis and the industrial flight hypothesis; some perspectives on theory and empirics. University of Oslo, Centre for Development and the Environment, Norway
- Keynes JM (1936) *The general theory of employment, interest and money*. Macmillan, London
- Kim TH, Lee YS, Newbold P (2004) Spurious nonlinear regressions in econometrics. School of Economics, University of Nottingham, Nottingham NG7 2RD, UK
- Koop G, Potter SM (2001) Are apparent findings of nonlinearity due to structural instability in economic timeseries? *Economet J* 4(1):37–55
- Kotchoni R (2018) Detecting and Measuring Nonlinearity. *Econometrics* 6(37):1–27
- Lau LS, Choong CK, Eng YK (2014) Investigation of the environmental Kuznets curve for carbon emissions in Malaysia: do foreign direct investment and trade matter? *Energy Policy* 68:490–497
- Levin A, Lin CF, Chu C-S (2002) Unit root in panel data: asymptotic and finite-sample properties. *J Econ* 108:1–24
- Liew VKS, Chong TTL, Lim KP (2003) The inadequacy of linear autoregressive model for real exchange rates: empirical evidence from Asian economies. *Appl Econ* 35:1387–1392
- Maddala GS, Wu S (1999) A comparative study of unit root tests with panel data and a new simple test. *Oxford Bull Econ Stat* 61:631–665
- Mensi W, Shahzad SJH, Hammoudeh S, Hamed A-YK (2017) Asymmetric impacts of public and private investments on the non-oil GDP of Saudi Arabia. *Int Econ* 156:1–16. <https://doi.org/10.1016/j.inteco.2017.10.003>

- Merican Y, Yusop Z, Noor ZM, Hook LS (2007) Foreign direct investment and the pollution in five ASEAN nations. *Int J Econ Manag* 1(2):245–261
- Mert M, Bölük G (2016) Do foreign direct investment and renewable energy consumption affect the CO<sub>2</sub> emissions? New evidence from a panel ARDL approach to Kyoto Annex countries. *Environ Sci Pollut Res* 23:21669–21681. <https://doi.org/10.1007/s11356-016-7413-7>
- Mert M, Caglar AE (2020) Testing pollution haven and pollution halo hypotheses for Turkey: a new perspective. *Environ Sci Pollut Res* 27:32933–32943. <https://doi.org/10.1007/s11356-020-09469-7>
- Mihajlović V, Marjanović G (2020) Asymmetries in effects of domestic inflation drivers in the Baltic states: a Phillips curve-based nonlinear ARDL approach. *Balt J Econ* 20(1):94–116. <https://doi.org/10.1080/1406099X.2020.1770946>
- Munir Q, Kok SC (2015) Purchasing power parity of ASEAN-5 countries revisited: heterogeneity, structural breaks and cross-sectional dependence. *Glob Econ Rev* 44(1):116–149. <https://doi.org/10.1080/1226508X.2015.1012091>
- Nguyen CP, Schinckus C, Su TD (2020) Economic integration and CO<sub>2</sub> emissions: evidence from emerging economies. *Clim Dev* 12(4):369–384. <https://doi.org/10.1080/17565529.2019.1630350>
- O'Connell PGJ (1998) The overvaluation of purchasing power parity. *J Int Econ* 44:1–20
- Öztürk Z, Öz D (2016) The Relationship between energy consumption, income, foreign direct investment, and CO<sub>2</sub> emissions: the case of Turkey. *Çankırı Karatekin Üniversitesi İİBF Dergisi* 6(2):269–288
- Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. *J Appl Econ* 22:265–312
- Pesaran MH, Smith R (1995) Estimating long-run relationships from dynamic heterogeneous panels. *J Econ* 68(1):79–113
- Pesaran MH, Shin Y, Smith RP (1999) Pooled mean group estimation of dynamic heterogeneous panels. *J Am Stat Assoc* 94:621–634. <https://doi.org/10.1080/01621459.1999.10474156>
- Pesaran MH, Shin Y, Smith RJ (2001) Bounds testing approaches to the analysis of level relationships. *J Appl Econ* 16:289–326
- Pesaran MH, Shin Y, Smith RP (2004) Pooled mean group estimation of dynamic heterogeneous panels. *ESE Discussion Papers* No: 16
- Pesaran MH (2004) General diagnostic tests for cross section dependence in panels. Working Paper No: 0435, University of Cambridge
- Rafindadi AA, Muye IM, Kaita RA (2018) The effects of FDI and energy consumption on environmental pollution in predominantly resource-based economies of the GCC. *Sustain Energy Technol Assess* 25:126–137
- Rahman UR, Chongbo W, Ahmad M (2019) An (a)symmetric analysis of the pollution haven hypothesis in the context of Pakistan: a non-linear approach. *Carbon Manag* 10(3):227–239. <https://doi.org/10.1080/17583004.2019.1577179>
- Sapkota P, Bastola U (2017) Foreign direct investment, income, and environmental pollution in developing countries: Panel data analysis of Latin America. *Energy Econ* 64:206–212
- Schorderet Y (2003) Asymmetric cointegration. Working paper No: 2003.01, University of Geneva
- Shahbaz M, Nasir MA, Roubaud D (2018) Environmental degradation in France: the effects of FDI, financial development, and energy innovations. *Energy Econ* 74:843–857
- Shin Y, Yu B, Greenwood-Nimmo M (2014) Modelling asymmetric cointegration and dynamic multipliers in a nonlinear ARDL framework. *Festschrift in honor of Peter Schmidt*. Springer, New York, pp 281–314
- Sichel DE (1993) Business cycle asymmetry: a deeper look. *Econ Inq* 31(2):224–236
- Solarin SA, Al-Mulali U, Musah I, Ozturk I (2017) Investigating the pollution haven hypothesis in Ghana: an empirical investigation. *Energy (Elsevier)* 124(C):706–719
- Sun C, Zhang F, Xu M (2017) Investigation of pollution haven hypothesis for China: an ARDL approach with breakpoint unit root tests. *J Clean Prod* 161:153–164



Tang CF, Tan BW (2015) The impact of energy consumption, income and foreign direct investment on carbon dioxide emissions in Vietnam. *Energy* 79:447–454

U.S. Energy information database. <https://www.eia.gov/>. Accessed 15 Sept 2020

Union of Concerned Scientists Data (2020) <https://www.ucsusa.org/resources/each-countrys-share-co2-emissions>. Accessed 1 Oct 2020

World Bank database. [www.worldbank.org](http://www.worldbank.org). Accessed 21 Sept 2020

Yavuz NÇ, Yilanci V (2012) Testing for nonlinearity in G7 macroeconomic time series. *Rom J Econ Forecast* 3:69–79

# An Investigation of Asymmetries in Exchange Rate Pass-Through to Domestic Prices



Fela Özbey

**Abstract** After the 2000–2001 financial crisis in Turkey, a strong reform program was initiated involving the enactment of the floating exchange rate regime in February 2001 and the adoption of inflation targeting as monetary policy in January 2002. This study aims to analyze the dynamics of exchange rate pass-through (ERPT) for the inflation-targeting period in Turkey by estimating the magnitudes of the short-run and long-run pass-through and by testing whether these magnitudes differ in contexts of depreciations and appreciations. To this end, the nonlinear autoregressive distributed lag (NARDL) approach is used. Since the bounds-testing procedure does not allow for stochastic seasonality or nonseasonal integration orders higher than one, to check the suitability of the series for this methodology, both seasonal and nonseasonal unit root tests are performed. The empirical results reveal asymmetry in the ERPT in both the short run and long run. In the long run, whereas appreciations of the domestic currency are not transmitted to domestic prices, the pass-through of depreciation is 43%. In the short run, the pass-through of appreciations is realized only in the current month and is 10.5%. The short-run pass-through of depreciations fluctuates over seven periods, and the total pass-through is approximately 3.5%.

**Keywords** Exchange Rate Pass-through · Nonlinear Autoregressive Distributed Lag Model · Asymmetric Level Relationships · Seasonal Unit Root Test

## 1 Introduction

ERPT is an important concept for inflation-targeting countries because nominal depreciations and appreciations are expected to affect inflationary pressures through cost and expectation channels.

Depreciations increase the prices of imported final goods and consequently the consumer price index (CPI), as the CPI basket consists mainly of traded goods. If

---

F. Özbey (✉)

Department of Econometrics, Faculty of Economics and Administrative Sciences, Çukurova University, Adana, Turkey  
e-mail: [fozbey@cu.edu.tr](mailto:fozbey@cu.edu.tr)

the dependence on imported intermediate goods is high and domestic substitutes for these goods are limited, an increase in the exchange rate increases production costs and consequently domestic prices. Therefore, the prices of imported intermediate goods and the prices of traded final goods are two possible cost channels through which exchange rates affect domestic prices. Additionally, since economic agents generally perceive the exchange rate as a nominal anchor and form their inflation expectations based on exchange rate movements, an increase in exchange rates is more likely to increase inflation expectations and thus the domestic inflation rate through the channel of expectations.

Although appreciations have weaker effects due to downward price stickiness, they are expected to alleviate inflationary pressures in the economy. Therefore, the exchange rate may emerge as a policy tool when monetary policy remains insufficient in fighting inflation by controlling interest rates. In this vein, a central bank must accurately forecast future movements in exchange rates and the magnitude of the ERPT on prices to set accessible targets for inflation. To develop more effective monetary policies in curtailing inflation, the dynamics of the pass-through should be analyzed and the length of time required for exchange rates to affect prices should be estimated.

A strong reform program involving the enactment of the floating exchange rate regime in February 2001 and the adoption of inflation targeting as monetary policy in January 2002 was initiated after the 2000–2001 financial crisis in Turkey. Although floating exchange rate regimes weaken the relationship between prices and exchange rates, the pass-through becomes partial but does not cease to exist. Additionally, because of the downward stickiness of prices, the pass-through is expected to be asymmetric. This study aims to analyze the dynamics of ERPT for the inflation-targeting period in Turkey by estimating the magnitudes of the short-run and long-run pass-through and by testing whether these magnitudes differ in contexts of depreciations and appreciations. To this end, the NARDL approach is used.

The NARDL approach is one of the methods used for modeling nonstationarity and nonlinearity jointly. This modeling approach proposes a simple nonlinear dynamic framework that is flexible enough to simultaneously embody asymmetries in the long-term and short-term relationships of  $I(1)$  and/or  $I(0)$  regressors and to employ testing procedures in the investigation of the stabilities and asymmetries of these relations.

## 2 Literature Review

One of the earliest studies examining the ERPT to Turkish domestic prices is by Leigh and Rossi (2002). Using a recursive vector autoregressive (VAR) model, they find that compared to the ERPT in other developing countries, the ERPT in Turkey is higher and completes in a shorter time; the effect of the exchange rate on prices is severe in the first four months and disappears a year later. Arat (2003) replicates the

VAR analysis of Leigh and Rossi by modifying the data and concludes that there is complete ERPT in the long run.

By using a threshold VAR model, Arbatlı (2003) investigates the asymmetries in ERPT to Turkish domestic prices for the period between January 1994 and May 2004. This study finds that the post-floating exchange rate period in Turkey is associated with a lower ERPT. It also finds that ERPT to Turkish domestic prices is asymmetric and that ERPT is lower during periods of output decreases, higher levels of inflation and depreciation, and considerable changes in exchange rates.

Kara and Ögünç (2005) perform a VAR analysis by separating their data into two subsets: before and after exchange rate regime-switching. Similar to Arbatlı (2003), they find that ERPT decreases after 2001.

Dinççağ (2009) examines whether the magnitude of ERPT differs during depreciations and appreciations using asymmetric cointegration analysis by separating their data into two subsets: before and after 2001. Similar to previous studies, she determines that the degree of ERPT is significantly decreased both in the short run and in the long run after 2001. Additionally, she detects that the ERPT is asymmetric and that the pass-through of depreciations is higher than appreciations.

Boz (2013), using a NARDL model, investigates the asymmetries of ERPT to prices for the inflation-targeting regime in Turkey in the period from January 2002 to October 2012. The results indicate that the ERPT is asymmetric, and in the short run, only the depreciations pass through on domestic prices. In the long run, the transmission of depreciations is higher than the transmission of appreciations.

Similar to Boz (2013), Karamelikli and Korkmaz (2016) investigate the ERPT on inflation for Turkey using the NARDL method and monthly data. They investigate the period between January 2003 and November 2015. They find that depreciations and appreciations asymmetrically increase prices in the short run, but in the long run, the effect of exchange rates on prices is symmetric and negative.

Çiftçi and Yılmaz (2018) employ the smooth transition regression model for the 2003–2017 period to investigate the nonlinear dynamics of inflation persistence and ERPT. The results show that inflation persistence and ERPT to CPI are higher in a regime with considerably high import price shocks, and for PPI, ERPT is more influential during a high depreciation regime.

### 3 Methodology

Over the past few decades, in light of the economic literature emphasizing the prevalence of nonlinearity and asymmetry in the behaviors of economic agents, econometric literature has provided significant momentum in modeling asymmetries. In addition, considering that most economic time series are generated by random walk processes, some studies attempt to model nonstationarity and nonlinearity jointly.

The NARDL approach of Shin et al. (2014) is one of the methods used for modeling nonstationarity and nonlinearity jointly. This modeling approach proposes a simple nonlinear dynamic framework that is flexible enough to simultaneously

embody asymmetries in the long-term and short-term relationships of I(1) and/or I(0) regressors and to employ testing procedures in the investigation of the stabilities and asymmetries of these relations.

The method implements the autoregressive distributed lag modeling approach of Pesaran and Shin (1998) using decomposed regressors as in Schorderet (2001), and it employs the bounds-testing procedure of Pesaran et al. (2001), hereafter PSS, to examine the existence of level relationships. In the presence of such relationships, long-term and short-term asymmetries are examined by testing for equality of the corresponding coefficients of the positive and negative components using the Wald test, named after Wald (1943).

The NARDL(p,q) model used in this study is in the following form:

$$LP_t = \sum_{i=1}^p \omega_i LP_{t-i} + \sum_{j=0}^q \left( \delta_j^+ LER_{t-j}^+ + \delta_j^- LER_{t-j}^- \right) + \varepsilon_t. \tag{1}$$

The linear error correction representation (ECR) of this NARDL model is

$$\begin{aligned} \Delta LP_t &= \gamma LP_{t-1} + \beta^+ LER_{t-1}^+ + \beta^- LER_{t-1}^- \\ &+ \sum_{i=1}^{p-1} \xi_i \Delta LP_{t-i} + \sum_{j=0}^{q-1} \left( \psi_j^+ \Delta LER_{t-j}^+ + \psi_j^- \Delta LER_{t-j}^- \right) + \varepsilon_t, \end{aligned} \tag{2}$$

and its nonlinear form is

$$\begin{aligned} \Delta LP_t &= \gamma \left( LP_{t-1} + \varphi^+ LER_{t-1}^+ + \varphi^- LER_{t-1}^- \right) \\ &+ \sum_{i=1}^{p-1} \xi_i \Delta LP_{t-i} + \sum_{j=0}^{q-1} \left( \psi_j^+ \Delta LER_{t-j}^+ + \psi_j^- \Delta LER_{t-j}^- \right) + \varepsilon_t. \end{aligned} \tag{3}$$

Here,  $\Delta$  is the difference operator,  $P_t$  is the CPI,  $ER_t$  is the average nominal exchange rate, and  $L$  is the natural logarithm.  $LER_t^+$  is the positive component of  $LER_t$  evaluated as  $LER_t^+ = \sum_{i=1}^t \max(\Delta LER_i, 0)$ ; and  $LER_t^-$  is the negative component of  $LER_t$  evaluated as  $LER_t^- = \sum_{i=1}^t \min(\Delta LER_i, 0)$ ;  $\varepsilon_t$  is a normally distributed white noise error term; and  $\omega_i, \delta_j^+, \delta_j^-, \gamma, \beta^+, \beta^-, \xi_i, \psi_j^+, \psi_j^-, \varphi^+$ , and  $\varphi^-$  are parameters. Here,  $\varphi^+ = \beta^+ / \gamma$  and  $\varphi^- = \beta^- / \gamma$ . A (restricted or unrestricted) time trend and/or a (restricted or unrestricted) constant may also be added to the model if needed.

Testing the existence of long-run relationships between prices and partial sums of changes in the nominal exchange rate can be performed with a bounds-testing procedure by using the Wald test statistic for the null hypothesis of  $\gamma = \beta^+ = \beta^- = 0$  in (2) or the t-test statistic for the null hypothesis of  $\gamma = 0$  against the alternative of  $\gamma < 0$  because of the nonlinear ECR in (3). The distributions of these statistics are nonstandard; therefore, critical values tabulated by Pesaran et al. (2001) are used.

Once the level relationships are detected, the long-run and short-run asymmetries can be tested by performing the Wald test for nulls of  $\beta^+ = \beta^-$  and  $\sum \psi_j^+ = \sum \psi_j^-$ , respectively.

Although the bounds-testing procedure does not require prior knowledge of the (trend or first difference) stationarity of regressors, the quarterly or monthly observed variables should be pretested for seasonal stochastic trends since the procedure does not allow models with explosive or seasonal unit roots. To test for stochastic seasonality, the testing procedure proposed by Hylleberg et al. (1990), hereafter HEGY, is preferred because it also allows testing for conventional (nonseasonal) unit roots in seasonal settings with the advantage of distinguishing processes that have unit roots only at particular seasonal frequencies.

The starting point of the HEGY test is the general autoregressive representation of the series  $y_t$ :

$$\phi(L)y_t = \varepsilon_t \tag{4}$$

where  $\phi(L)$  is the lag operator polynomial and  $\varepsilon_t \sim iid(0, \sigma^2)$ . The frequency of a root is the value of  $\theta$  in its polar representation  $e^{i\theta}$ . Seasonal roots correspond to  $\theta = 2\pi j/M$ , where  $j = 1, 2, \dots, M - 1$ , and  $M$  is the number of observation periods in a year.

The testing procedure linearizes  $\phi(L)$  around the nonseasonal (zero frequency) and  $M - 1$  seasonal unit roots by expressing the autoregressive polynomial in terms of the sum of elementary polynomials and a remainder:

$$\phi(L) = \sum_{k=1}^M \alpha_k \Delta(L) / \delta_k(L) + \Delta(L) \tilde{\phi}(L). \tag{5}$$

Representing the multiplication of  $y_t$  with the corresponding elementary polynomial by  $x_{k,t}$  and  $\Delta(L)y_t$  by  $x_{13,t}$ , the reorganized process for monthly data is expressed as

$$\tilde{\phi}(L)x_{13,t} = \sum_{k=1}^{12} \alpha_k x_{k,t-1} + \varepsilon_t. \tag{6}$$

Deterministic components as a time trend, a constant, and seasonal dummies may also be added to the model if needed. Thus, the testing procedure turns to estimating (6) and testing various hypotheses related to the coefficients of the model.

To test for unit roots at zero and  $\pi$  frequencies, the nulls of  $\alpha_k = 0$  against the alternatives  $\alpha_k < 0$ <sup>1</sup> should be examined using relevant t-statistics. For frequencies other than zero and  $\pi$ , the joint null hypothesis of  $\alpha_{k-1} = \alpha_k = 0$  should be examined using F-statistics.

---

<sup>1</sup> For coefficient estimates greater than one, the two-sided alternative of nonunity (which allows for testing the existence of an explosive root) should be used.

Another limitation of the bounds-testing procedure is that it does not allow for the I(2) series. To check the number of nonseasonal stochastic trends, the widely used augmented Dickey-Fuller (ADF) stationarity test is also performed.

The ADF test of Dickey and Fuller (1981) suggests modeling a time series as an AR(p) model and testing for the presence of a unit root. The testing procedure exploits the fact that an AR model is a difference equation with a (trend) stationary forcing process, so the stability of the AR process guarantees the absence of a stochastic trend. Therefore, if the series under consideration has deterministic components, these components should also be modeled to eliminate their effects on the distributions of the test statistics. To capture the possible (nonseasonal) deterministic components, Dickey and Fuller (1979) suggested adding a constant or/and a deterministic trend to the model if necessary. The final version of the models used in the ADF testing procedure is as follows:

$$\Delta y_t = c + \rho y_{t-1} + \varpi t + \sum_{i=1}^{p-1} \eta_i \Delta y_{t-i} + \varepsilon_t \quad (7)$$

$$\Delta y_t = c + \rho y_{t-1} + \sum_{i=1}^{p-1} \eta_i \Delta y_{t-i} + \varepsilon_t \quad (8)$$

$$\Delta y_t = \rho y_{t-1} + \sum_{i=1}^{p-1} \eta_i \Delta y_{t-i} + \varepsilon_t. \quad (9)$$

Thus, the test for the presence of a stochastic trend (i.e., unit root) turns to a test for significance of  $\rho$  in (7), (8), and (9) against (mostly) the one-sided alternative. The presence of the deterministic component can also be tested using various joint hypotheses. Critical values for testing the presence of a unit root first are simulated in Dickey and Fuller (1979) and are updated in MacKinnon (1996). Critical values for testing joint hypotheses are given in Dickey and Fuller (1981). Considering the properties of ADF test statistics, an appropriate testing strategy is proposed in Dolado et al. (1990).

## 4 Testing for Asymmetries in ERPT

This study uses monthly data of the CPI in 2010 prices and the average nominal exchange rates of US dollars and euros for the period of 2002:01–2020:11. Data are retrieved from the IFS database of the International Monetary Fund (2021).

Because lag lengths suggested by the Bayesian information criterion (BIC) do not provide white noise errors for some models, to avoid the autocorrelation problem, this study follows Stock and Watson (2019, pp.538) and determines lag lengths for all models in testing and estimation procedures using the Akaike information criterion (AIC). If the lag length suggested by the AIC is too long, the lags are reduced one by one until the shortest lag that guarantees the uncorrelated errors is achieved, and this user-specified lag is used.

As noted previously, the bounds-testing procedure is not applicable for series with seasonal stochastic trends. Because the frequency of the data used in this study is monthly, the HEGY test is applied to test whether series have (seasonal and nonseasonal) stochastic trends and to check the appropriateness of the variables for this method. The test results for the  $LP_t$  and  $LER_t$  series are given in Table 1 and Table 2, respectively.

Although the HEGY test results reject the presence of a seasonal root in all models used, there is no agreement in the results for zero-frequency unit roots. To determine the number of nonseasonal stochastic trends present in the series, the ADF test is also performed by using the testing strategy of Dolado et al. (1990). The results of the test are given in Table 3.

According to the Dolado et al. (1990) strategy for performing the ADF test, the test concludes using the model with the intercept and critical values of the standard normal distribution. Because the bounds-testing procedure is not applicable for series generated by explosive processes, the two-sided alternative of nonunity (which allows for testing the existence of an explosive root) is also tested. The results for  $LP_t$  in both cases are in favor of the unit root. For the series  $LER_t$ , the null of the unit root cannot be rejected against the one-sided alternative of stability but against the two-sided alternative of nonunity, the null can be rejected in favor of the explosive root at only the 10% significance level. For the differenced series, the ADF testing procedure starts with the model with the intercept only because differenced series do not contain trends. For both series, the null of the unit root is rejected at the 1% significance level. In conclusion, at the 5% significance level, both series are integrated of order 1.

**Table 1** HEGY test results for  $LP_t$

Frequency	With Seasonal Dummies			Without Seasonal Dummies		
	Trend and Constant	Constant	None	Trend and Constant	Constant	None
0	-0.061	-1.806	-1.806	-0.302	-1.966	-4.235***
$2\pi / 12$ and $22\pi / 12$	17.441***	17.902***	17.902***	11.430**	12.012**	11.998**
$4\pi / 12$ and $20\pi / 12$	13.930***	14.043***	14.043***	5.379*	5.462*	5.555*
$6\pi / 12$ and $18\pi / 12$	22.585***	22.793***	22.793***	12.505**	12.680**	12.823**
$8\pi / 12$ and $16\pi / 12$	17.478***	17.623***	17.623***	8.844**	8.959**	9.038**
$10\pi / 12$ and $14\pi / 12$	26.675***	26.884***	26.884***	20.199**	20.417**	20.602**
$\pi$	-4.193***	-4.207***	-4.207***	-3.966***	-3.989***	-4.008***
All seasonal	24.363***	25.107***	25.107***	14.847**	15.382**	15.556**
All frequencies	22.782***	23.801***	23.801***	13.974**	14.713**	15.973**
Lag (AIC)	1	1	1	1	1	1

\*\*\*, \*\*, and \* indicate rejection of the null hypothesis at the 1%, 5%, and 10% significance levels, respectively



**Table 2** HEGY test results for  $LER_t$

Frequency	With Seasonal Dummies			Without Seasonal Dummies		
	Trend and Constant	Constant	None	Trend and Constant	Constant	None
0	-0.150	-4.106***	-4.106***	-0.398	-3.747***	-4.650***
$2\pi / 12$ and $22\pi / 12$	23.402***	19.579***	19.579***	24.627**	25.672**	25.331**
$4\pi / 12$ and $20\pi / 12$	14.165***	12.729***	12.729***	11.055**	11.307**	11.589**
$6\pi / 12$ and $18\pi / 12$	11.993***	9.900***	9.900***	13.246**	13.381**	13.526**
$8\pi / 12$ and $16\pi / 12$	21.650***	18.092***	18.092***	20.139**	20.186**	20.123**
$10\pi / 12$ and $14\pi / 12$	28.979***	27.577***	27.577***	18.836**	19.006**	19.147**
$\pi$	-5.728***	-3.820***	-3.820***	-4.533***	-4.566***	-4.621***
All seasonal	27.766***	27.028***	27.028***	23.820**	24.427**	24.527**
All frequencies	25.541***	26.236***	26.236***	21.877**	24.358**	25.252***
Lag (AIC)	3	6	6	4	4	4

\*\*\*, \*\*, and \* indicate rejection of the null hypothesis at the 1%, 5%, and 10% significance levels, respectively

**Table 3** ADF test results for levels and first differences of  $LP_t$  and  $LER_t$

Series	Model	Lag	$\tau$ statistic	$\phi$ statistic	z statistic
$LP_t$	Trend and intercept	6	-1.844581	1.778647	
	Intercept	6	0.220932	8.247594***	0.220932
$\Delta LP_t$	Intercept	5	-4.594879***		
$LER_t$	Trend and intercept	2	-0.261601	2.083593	
	Intercept	2	1.727160	6.376468**	1.727160 <sup>2</sup>
$\Delta LER_t$	Intercept	1	-10.89087***		

\*\*\*, \*\*, and \* indicate rejection of the null hypothesis at the 1%, 5%, and 10% significance levels, respectively

<sup>2</sup> For the one-sided alternative, the null hypothesis cannot be rejected at any significance level, for the two-sided alternative, the null can be rejected only at the 10% significance level.

For all model types used in the HEGY test, the results of the hypothesis tests show that all seasonal roots are smaller than unity in magnitude; that is, the logarithmic transformations of the  $P_t$  and  $ER_t$  series do not contain any seasonal stochastic trends, and the ADF test results indicate that both series are I (1) over the observed period. Therefore, the bounds-testing approach is applicable for these series.

To model and test the asymmetries in the relationships between prices and exchange rates, in the first step, the  $LER_t$  series is decomposed into its negative and positive components as  $LER_t = LER_t^- + LER_t^+$ .

The deterministic components of the NARDL model are set up as restricted intercept and no trend (Case II of PSS). According to the AIC, the optimal model is selected as NARDL(10,7,1), but standardized residuals of this model reveal two outliers that are dated as 2011M6 and 2018M9.<sup>3</sup> Assigning two dummy variables ( $D_{2011:06}$  and  $D_{2018:09}$ ) to these dates, the appropriate NARDL model with normally distributed white noise residuals is selected as NARDL(8,7,1). The stability of the model is checked by using cumulative sum (CUSUM) and cumulative sum of squares (CUSUMSQ) statistics suggested by Brown et al. (1975). Parameter estimates and diagnostic test results for this model are given in Table 4, and CUSUM and CUSUMSQ statistics graphs are given in Fig. 1.

As diagnostic tests, Ljung-Box Q-statistics for the 1st, 4th, and 12th lags of the residuals are used to examine the autocorrelation problem, the White test is used to examine the heteroscedasticity problem, the Ramsey RESET is used to examine the existence of omitted variables, and the Jarque-Bera test is used to examine the normality of the error term in the model. None of the nulls of these tests can be rejected. These results demonstrate that the residuals are generated by a normally distributed white noise process and that the model does not suffer from a functional form problem.

The CUSUM and CUSUMSQ statistics lie inside the 5% confidence interval, indicating stability in the equation parameters and the variance during the sample period.

Shin et al. (2014) point out that because of the dependence between positive and negative components, the results of bounds tests are more conservative when these components are counted as a single variable. The bounds test is performed by taking into account this notification. The test result is in favor of the long-run level relationship between  $LP_t$  and the components of  $LER_t$  at the 1% significance level.

Because all variables are in logarithms, the coefficients of the exchange rate components indicate the elasticities of the price index with respect to exchange rates; hence, they indicate the magnitude of ERPT. The long-run coefficient of the appreciations is statistically insignificant. This shows that because of the downward stickiness of the prices, in the long run, the appreciations of the domestic currency do not pass through to the prices. The pass-through of depreciations in the long run is statistically significant and is approximately 42.7%. The speed of the adjustment coefficient is -0.064, indicating that when other factors are held constant, the adjustment to the long-run equilibrium is completed in 16 months.

Since the long-run depreciation coefficient is statistically significant while the long-run appreciation coefficient is statistically insignificant, it can be concluded that there is asymmetry in the ERPT in the long run without conducting the Wald test for the equity of the long-run coefficients.

According to the results, the pass-through of appreciations is realized only in the current month and is approximately 10.5%. The pass-through of depreciations in the

---

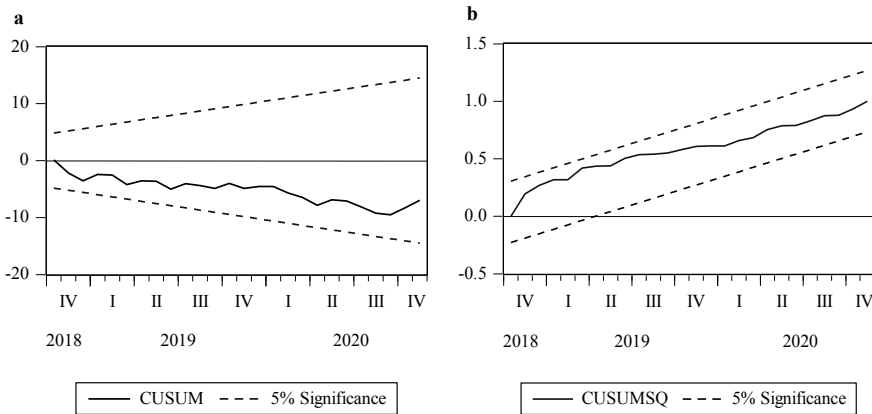
<sup>3</sup> Standardized residuals greater than 3 in magnitude.

**Table 4** Coefficients estimates and diagnostic tests of NARDL (8, 7, 1) ECR

Linear ECR		Long-Run Equation in Restricted ECR	
Variable	Coefficient	Variable	Coefficient
Constant	0.253407***	Constant	3.960753***
$LP_{t-1}$	-0.063980*** <sup>4</sup>	$LER_{t-1}^+$	0.426840***
$LER_{t-1}^+$	0.027309***	$LER_{t-1}^-$	-0.094538
$LER_{t-1}^-$	-0.006048		
$\Delta LP_{t-1}$	0.448472***		
$\Delta LP_{t-2}$	-0.167845**		
$\Delta LP_{t-3}$	0.118003*		
$\Delta LP_{t-4}$	-0.177437***		
$\Delta LP_{t-5}$	-0.002636	Diagonistics	
$\Delta LP_{t-6}$	0.231335***	$R^2$	0.999830
$\Delta LP_{t-7}$	-0.160048***	$Adj R^2$	0.999813
$\Delta LER_t^+$	0.096719***	$Q(1)$	0.5051
$\Delta LER_{t-1}^+$	-0.044526**	$Q(4)$	1.2683
$\Delta LER_{t-2}^+$	0.046227**	$Q(12)$	10.677
$\Delta LER_{t-3}^+$	-0.053697***	$F_{WH}$	0.859276
$\Delta LER_{t-4}^+$	0.035899*	$F_{RR}$	0.454643
$\Delta LER_{t-5}^+$	0.025351	$\chi_{JB}^2$	3.053221
$\Delta LER_{t-6}^+$	-0.045784**	$F_{PSS}$	10.71551***
$\Delta LER_t^-$	0.104631***	$F_{WSR(1)}$	0.626858
$D_{2011:06}$	-0.027004***	$F_{WSR(2)}$	1.526288
$D_{2018:09}$	0.040867***	$F_{WSR(3)}$	4.155571**

\*\*\*, \*\*, and \* indicate rejection of the null hypothesis at the 1%, 5%, and 10% significance levels, respectively.  $Q(1)$ ,  $Q(4)$ , and  $Q(12)$  are Ljung-Box Q-statistics for the 1st, 4th, and 12th lags, respectively;  $F_{WH}$  is the White heteroscedasticity test statistic;  $F_{RR}$  is the functional form Ramsey RESET test statistic;  $\chi_{JB}^2$  is the Jarque–Bera normality test statistic;  $F_{PSS}$  is the Pesaran-Shin-Smith test statistic for testing level relationships;  $F_{WSR(1)}$  is the Wald test statistic for testing short-run asymmetry using all short-run parameter estimates;  $F_{WSR(2)}$  is the Wald test statistic for testing short-run asymmetry using short-run coefficients that are statistically significant at the 10% significance level;  $F_{WSR(3)}$  is the Wald test statistic for testing the short-run asymmetry using short-run coefficients that are statistically significant at the 5% significance level

<sup>4</sup> Asterisks of this coefficient are based on the  $F_{PSS}$  test result.



**Fig. 1** a CUSUM and b CUSUMSQ graphs

short run fluctuates over seven periods. The total short-run pass-through of depreciations (evaluated as the sum of the statistically significant short-run coefficients of depreciations) is equal to 3.5%.

To test for asymmetry in the short-run ERPT, three different hypotheses are constructed. The test statistics of these tests are reported in Table 4 as  $F_{WSR(1)}$ ,  $F_{WSR(2)}$ , and  $F_{WSR(3)}$ . First, short-run symmetry is tested by using all parameter estimates.  $F_{WSR(1)}$  is the Wald test statistic related to this hypothesis test. The null hypothesis of this test cannot be rejected. The second hypothesis test for short-run symmetry is constructed by using only coefficients that are statistically significant at the 10% significance level, that is, by excluding the coefficient of the fifth lag of the  $LER_t^+$ . The Wald test statistic denoted by  $F_{WSR(2)}$  relates to this test. The null hypothesis of this test also cannot be rejected. The third and last hypothesis test for short-run symmetry is constructed by using coefficients that are statistically significant at the 5% significance level. The test statistic of this hypothesis test is reported as  $F_{WSR(3)}$ . The null hypothesis of this test is rejected at the 5% significance level, indicating short-run asymmetry in the ERPT to domestic prices.

## 5 Conclusion

After the 2000–2001 financial crisis in Turkey, a strong reform program was initiated involving the enactment of the floating exchange rate regime in February 2001 and the adoption of inflation targeting as monetary policy in January 2002. In the floating exchange rate regime, the relationship between exchange rates and prices is weaker than that in other exchange rate regimes. Additionally, although most studies in the economic and econometric literature model the ERPT symmetrically, the downward stickiness of prices is expected to destroy the symmetry of the pass-through. This

study aims to analyze the ERPT dynamics for the inflation-targeting period in Turkey by estimating the magnitudes of the long run and short run pass-through and by testing whether these magnitudes differ during depreciations and appreciations.

Although the NARDL approach is a relatively flexible methodology that models nonstationarity and nonlinearity jointly, it has some limitations. The bounds-testing procedure for testing long-run level relationships does not allow for stochastic seasonality or nonseasonal integration orders higher than one. To check the suitability of the series for this methodology, both seasonal and nonseasonal unit root tests are performed, and both series are detected to be  $I(1)$  and suitable.

The results show that there is asymmetry in the ERPT in both the short run and the long run. In the long run, appreciations of the domestic currency do not pass through to domestic prices. This can be explained by the downward stickiness of prices. The pass-through of depreciations in the long run is approximately 43%. The speed of the adjustment coefficient is  $-0.064$ , indicating that with other factors held constant, the adjustment to the long-run equilibrium is completed in 16 months.

According to the results, the pass-through of appreciations is realized only in the current month and is approximately 10.5%. The pass-through of depreciations in the short run fluctuates over seven periods. The total short-run pass-through of depreciations (evaluated as the sum of the statistically significant short-run coefficients of depreciations) is approximately 3.5%.

**Acknowledgements** A preliminary version of this study entitled “An Investigation of Exchange Rate Pass-through in Turkey” was presented at the 2016 WEI International Academic Conference, Rome, Italy. This research was supported by the Scientific Research Projects Unit of Çukurova University Grant SED-2016-7705.

## References

- Arat K (2003) Türkiye’de Optimum Döviz Kuru Rejimi Seçimi ve Döviz Kurlarından Fiyatlara Geçiş Etkisinin İncelenmesi. Expert Thesis, Central Bank of Turkey, Ankara, Turkey
- Arbatlı EC (2003) Exchange rate pass-through in Turkey: looking for asymmetries. *Central Bank Rev* 3(2):85–124
- Boz Ç (2013) Asymmetric ERPT for the periods of implicit-explicit IT in Turkey. *Scott J Arts, Soc Sci Stud* 9(2):3–11
- Brown RL, Durbin J, Evans JM (1975) Techniques for testing the constancy of regression relationships over time. *J Roy Stat Soc Ser B (Methodol)* 149–163
- Çiftçi M, Yılmaz MH (2018) Nonlinear dynamics in exchange rate pass-through and inflation persistence: The case of Turkish economy. *Asian J Econ Model* 6(1):8–20
- Dickey DA, Fuller WA (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49(4):1057–1072
- Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc* 74(366a):427–431
- Diñççağ A (2009) Exchange rate pass-through in Turkey: asymmetric cointegration analysis. *Bilkent University*
- Dolado JJ, Jenkinson T, Sosvilla-Rivero S (1990) Cointegration and unit roots. *Journal of Economic Surveys* 4(3):249–273

- Hylleberg S, Engle RF, Granger CW, Yoo BS (1990) Seasonal integration and cointegration. *Journal of Econometrics* 44(1–2):215–238
- International Monetary Fund (2021) *International Financial Statistics (IFS)*. Accessed 01/20/2021
- Kara H, Ögünç F (2005) Exchange rate pass-through in Turkey: it is slow, but is it really low? Central bank of the republic of Turkey. Research department working paper No. 05/10
- Karamelikli H, Korkmaz S (2016) The dynamics of exchange rate pass-through to domestic prices in Turkey. *Journal of Business Economics and Finance* 5(1):39–48
- Leigh D, Rossi M (2002) Exchange rate pass-through in Turkey. IMF Working Paper No. 02/204
- MacKinnon JG (1996) Numerical distribution functions for unit root and cointegration tests. *J Appl Economet* 11(6):601–618
- Pesaran MH, Shin Y (1999) An autoregressive distributed-lag modelling approach to cointegration analysis. In: Strøm S (ed) *Econometrics and economic theory: the Ragnar Frisch centennial symposium*. Cambridge University Press, Cambridge, pp 371–414
- Pesaran MH, Shin Y, Smith RJ (2001) Bounds testing approaches to the analysis of level relationships. *J Appl Economet* 16(3):289–326. <https://doi.org/10.1002/jae.616>
- Schorderet Y (2001) Revisiting Okun’s law: an hysteretic perspective. University of California, San Diego, Discussion Paper No.2001–13
- Shin Y, Yu B, Greenwood-Nimmo M (2014) Modelling Asymmetric Cointegration and Dynamic Multipliers in a Nonlinear ARDL Framework. In: C.Sickles R, C.Horrace W (eds) *Festschrift in Honor of Peter Schmidt Econometric Methods and Applications*. Springer, New York, pp 281–314
- Stock JH, Watson MW (2019) *Introduction to econometrics*. 4th edn. Pearson
- Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc* 54(3):426–482

# Investigation of the Country-Specific Factors for URAP



Eda Yalçın Kayacan and Aygül Anavatan

**Abstract** The international rankings of universities have a significant impact on the perception of academics, students, governments, and businesses toward the universities. The aim of the study is to identify country-specific factors that are thought to influence academic performance and to reveal how the country-specific factors should be developed for universities to increase their success. For this purpose, the University Ranking by Academic Performance index for universities in 103 countries covering the period 2013–2019 was taken as the dependent variable to represent academic performance. The index value, which is used as the dependent variable, was constituted by taking the average of the countries to which entered the ranking universities belong. The country-specific factors were taken as the political stability and absence of violence/terrorism, the rule of law, the freedom of the press, the economic freedom index, the university-industry collaboration in research & development index, and the gross domestic product per capita. The factors affecting the academic performance of universities were analyzed with spatial panel data methods, and the findings revealed that the rule of law, the university-industry collaboration, and the GDP per capita increase academic performance.

**Keywords** University ranking · URAP · Spatial panel data

## 1 Introduction

The international rankings of universities have a significant impact on the perception of academics, students, governments, and businesses toward the universities. The rankings ensure a quantitative comparison of educational institutions. Data to be

---

E. Y. Kayacan

Department of Statistics, Faculty of Science and Literature, Pamukkale University, Denizli, Turkey  
e-mail: [eyalcin@pau.edu.tr](mailto:eyalcin@pau.edu.tr)

A. Anavatan (✉)

Department of Econometrics, Faculty of Economics and Administrative Sciences, Pamukkale University, Denizli, Turkey  
e-mail: [aanavatan@pau.edu.tr](mailto:aanavatan@pau.edu.tr)

used in university rankings can be obtained from surveys, independent third parties, or university sources. Nevertheless, there is no consensus among the authors on the definition of quality and what an ideal indicator should be. So, the rankings are widely criticized because of the unsatisfactory criteria which they employ. Many rankings have revealed in the world over time in order to recruit the others' deficiencies.

The *US News and World Report*'s 1983 ranking for US universities is a milestone in the ranking system. After that, the first global university ranking was released by SJTUIHE (Shanghai Jiao Tong University Institute of Higher Education—China) in 2003. In addition to international rankings such as ARWU (Academic Ranking of World Universities—China), QS (Quacquarelli Symonds—United Kingdom), THE (Times Higher Education—United Kingdom), CHE (Centre for Higher Education Development—Germany), CWTS (*Centrum voor Wetenschap en Technologische Studies*—the Netherlands) Leiden Ranking, CWUR (Center for World University Rankings—United Arab Emirates), HEEACT (Higher Education Evaluation and Accreditation Council of Taiwan) or NTU (National Taiwan University) Ranking, U-Multirank (Germany), and Webometrics (Spain), national rankings have also been made in recent years. University Ranking by Academic Performance (URAP) has provided both world and national rankings since 2010 in Turkey. Calculating the URAP index used the six indicators which have the weights by 21%, 21%, 10%, 18%, 15%, and 15%, respectively: the number of articles, citations, total documents, journal impact total, journal citation impact total, and international collaboration (URAP). Although the number of universities is 400 in THE, 500 in ARWU, and around 700 in QS, URAP is a highly comprehensive ranking from the point of covering about 2500 universities.

In the literature, there exist numerous studies (Yonezawa et al. (2002), Liu and Liu (2005), Usher and Savino (2007), Saka and Yaman (2011), Boulton (2011), Rauhvargers (2013), Marginson (2014), Alaşehir et al. (2014), Mori (2016), Hammarfelt et al. (2017), and McAleer et al. (2019)) that introduce ranking systems. Kivinen and Hedman (2008) and Huang (2012) indicated some shortcomings in rankings. Several studies such as Marginson (2007), Çakır et al. (2015), and Moed (2017) compared ranking systems. Alaşehir (2010), Goglio (2016), Johnes (2018), and Uslu (2018) suggested alternative methodologies to evaluate academic performance. Teichler (2011), Arimoto (2011), and Elken et al. (2016) focused on how universities were transformed because of rankings. Marginson (2009), Ishikawa (2009), and Kehm (2014) criticized the ranking system becoming an emerging hegemony and side effects on the organizational behavior of universities. Pusser and Marginson (2013) discussed the role of appearing of state, social, and university power on the university rankings.

Grewal et al. (2008) fitted a logit model with data from U.S. News and World Report for the period 1999–2006 by including lagged rank to understand the competition for ranking. Frenken et al. (2017) analyzed the factors underlying research performance with the context of research excellence, internationalization, and innovation by applying OLS regression. Mingers et al. (2017) explored the extent to which several citation-based metrics from Google Scholar (GS) for all 130 UK



universities could be used to construct for the evaluation of the universities' performance. Pietrucha (2018) investigated country-specific factors that affect ARWU, QS, and THE by running cross-sectional regression. Meseguer-Martinez et al. (2019) examined the linkage between the uploaded YouTube videos and university ranking. Sebetci and Aksu (2019) researched the similarities and differences of universities in Turkey by using the URAP-TR index via the Multiple Correspondence Analysis Method.

In this study, we investigate whether the macroeconomic factors affect the URAP global ranking measuring academic performance. In other words, we focus on the question: do the economic, political, and socio-cultural conditions of a country affect the academic score of the universities in that country? The reason why we give preference to the URAP index in our study is to handle as many countries as possible. While evaluating country-specific factors in the study, the fact that the number of countries studied is comprehensive, and the use of spatial panel data techniques are the contributions to the related literature. Also, we readdress the country-specific factors by considering the variables political stability, the rule of law, freedom of the press, economic freedom index, university-industry collaboration, and GDP per capita. We reached the conclusion that the rule of law, university-industry collaboration, and GDP per capita increase academic performance. The rest of the study is organized as follows. The spatial panel data techniques are explained in the methodology. After that, the data used in the analysis is introduced and the empirical results are presented. Finally, we conclude in the last section.

## 2 Methodology

### 2.1 *Spatial Effects, Spatial Dependence, and Spatial Weight Matrix*

Spatial effects consist of two parts such as spatial autocorrelation, which expresses the interrelatedness of the observations due to the proximity between the places where the observations are made, and spatial heterogeneity, which refers to inhomogeneous parameters that change according to the location of the space. Spatial econometric methods enable model estimates that take into account the spatial autocorrelation and spatial heterogeneity effects seen in regression models estimated with cross-section and panel data (Anselin, 2001; Gerkman, 2010; Gülel, 2018).

In the spatial regression model estimation, the spatial weight matrix showing the neighborhood relationship between the spaces is used to include spatial effects in the model. In obtaining the weight matrix, four different weighting methods are preferred, geographical weighting, socioeconomic weighting, border neighborhood, and distance-based neighborhood (Gürel-Günel and Altuğ, 2016).

It is possible to pass the estimation of econometric models including spatial dependence after tests showing the presence of spatial dependence is performed. In the test

of spatial dependence, the neighborhood ratio statistic developed by Geary (1954) or the Moran I test statistic, which was developed by Cliff and Ord (1973), is used.

### 2.2 Spatial Panel Models

There are some points to be considered in the effective selection of components in spatial panel models. Choosing the appropriate model should be started by considering the model with all specific effects and all spatial components. First, all spatial components in the model should be preserved and reduced to the specific effective model as fixed effects or random effects. Then the spatial components should be reduced in the model established with specific effects. It is possible to control this process by doing the same in reverse. In other words, the selected model in the process should be compared with the final model, which is achieved by first reducing its spatial components and then its specific effects (Kopczewska et al., 2017). Therefore, there are two important tests in predicting spatial panel models. The first of these tests is the Hausman Test, which decides the panel data model type by choosing between the fixed effect and random effect panel data model. The other is the LM test, which determines the structure of spatial dependence and the appropriate spatial model (Anselin et al., 2008).

A general panel model that includes both, all specific effects (fixed effects and random effects) and all spatial coefficients, is expressed as in Eq. (1).  $\tau = 0$  and  $\tau \neq 0$  indicate the static model and the dynamic model, respectively, in Eq. (1) (Belotti et al., 2017).

$$y_{it} = \alpha + \tau y_{it-1} + \rho \sum_{j=1}^n w_{ij} y_{jt} + \sum_{k=1}^K x_{itk} \beta_k + \sum_{k=1}^K \sum_{j=1}^n w_{ij} x_{jtk} \theta_k + \mu_i + \gamma_t + v_{it} \tag{1}$$

$$v_{it} = \lambda \sum_{j=1}^n w_{ij} v_{jt} + \epsilon_{it}, \text{ for } i = 1, \dots, n \text{ and } t = 1, \dots, T$$

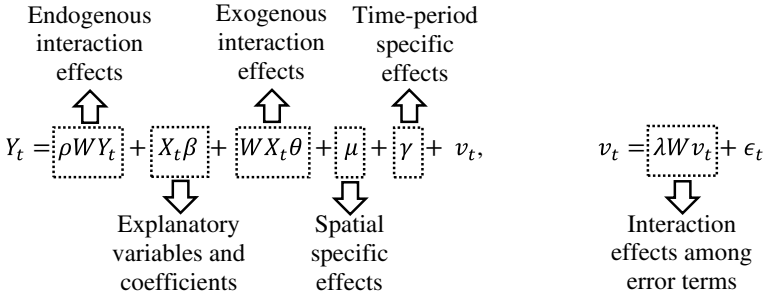
In Eq. (1),  $\rho$ ,  $\lambda$ , and  $\theta$  are, respectively, spatial autoregressive coefficient, spatial autocorrelation coefficient, and spatial Durbin coefficient.

There are many interaction effects such as internal, external, explanatory variables, time, and space in Eq. (1). Figure 1 demonstrates all these interaction effects on a general model expressing in matrix notation.

$$Y_t = \rho W Y_t + X_t \beta + W X_t \theta + \mu + \gamma + v_t, v_t = \lambda W v_t + \epsilon_t$$

There are three methods used to predict models in which interaction effects are included. These methods are maximum likelihood or quasi-maximum likelihood, instrumental variables or generalized method of moments, and Bayesian Markov Chain Monte Carlo approach (Elhorst, 2011).

Different models can be reached by putting a restraint on the coefficients in Eq. 1. If  $\lambda = 0$  and  $\theta = 0$ , the Spatial Autoregressive Model (SAR) emerges. Considering the specific effects, the spatial autoregressive model with fixed effect is expressed



**Fig. 1** Interaction effects in a spatial panel model (Bi et al., 2018)

by Eq. 2, and the spatial autoregressive model with random effect is represented by Eq. 3.

$$y_{it} = \rho \sum_{i \neq j}^n w_{ij} y_{jt} + x_{it} \beta + \mu_i + v_{it}, v_{it} \sim N(0, \sigma^2) \tag{2}$$

$$y_{it} = \rho \sum_{i \neq j}^n w_{ij} y_{jt} + x_{it} \beta + \mu + v_{it}, v_{it} = \mu_i + \epsilon_{it}, \epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{3}$$

If  $\rho = 0$  and  $\theta = 0$  in Eq. (1), the general spatial panel model turns into the Spatial Error Model (SEM). Considering the specific effects, it is expressed as the spatial error model with fixed effect as follows:

$$y_{it} = x_{it} \beta + \mu_i + v_{it}, v_{it} = \lambda \sum_{i \neq j}^n w_{ij} v_{jt} + \epsilon_{it}, \epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{4}$$

In the literature, there are two SEM specifications which are SEM-RE (spatial error model with random effect) and RE-SEM (random effect with spatial error model). The models of SEM-RE and RE-SEM are, respectively, represented by Eqs. 5 and 6 (Salima et al., 2018).

$$y_{it} = x_{it} \beta + v_{it}, v_{it} = \mu_i + \rho \sum_{i \neq j}^n w_{ij} v_{jt} + \epsilon_{it}, \epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{5}$$

$$y_{it} = x_{it} \beta + \mu + v_{it}, v_{it} = \rho \sum_{i \neq j}^n w_{ij} v_{jt} + \epsilon_{it}, \epsilon_{it} = \mu_i + \epsilon_{it}, \epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma^2) \tag{6}$$

If  $\lambda = 0$  in Eq. (1), the Spatial Durbin Model (SDM) arises. Spatial Durbin Model (SDM) with the fixed effects is indicated as follows:

$$y_{it} = \rho \sum_{i \neq j}^n w_{ij} y_{jt} + x_{it} \beta + \sum_{i \neq j}^n w_{ij} x_{jt} \theta + \mu_i + v_{it}, v_{it} \sim N(0, \sigma^2) \quad (7)$$

If  $\theta = 0$  in Eq. (1), the spatial autoregressive model with autoregressive disturbances (SAC) or the other name spatial autoregressive model with autoregressive disturbances (SARAR) is obtained. The SAC model with the fixed effects is as in Eq. (8).

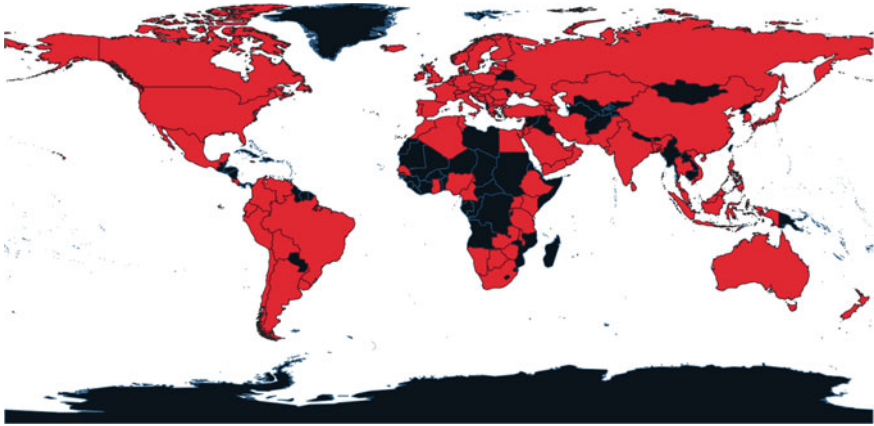
$$y_{it} = \rho \sum_{i \neq j}^n w_{ij} y_{jt} + x_{it} \beta + \mu_i + v_{it}, v_{it} = \lambda \sum_{i \neq j}^n w_{ij} v_{it} + \epsilon_{it}, \epsilon_{it} \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad (8)$$

Finally, if  $\rho = 0, \theta = 0$ , and  $\mu_i = \phi \sum_{j=1}^n w_{ij} \mu_j + \eta_i$  in Eq. (1), the following generalized spatial panel random effects model (GSPRE) emerges:

$$y_{it} = x_{it} \beta + \mu_i + v_{it}, v_{it} = \lambda \sum_{i \neq j}^n w_{ij} v_{it} + \epsilon_{it}, \mu_i = \phi \sum_{j=1}^n w_{ij} \mu_j + \eta_i, \epsilon_{it} \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad (9)$$

### 3 Dataset and Empirical Results

The dataset covering the 2013–2019 time-period includes the URAP ranking values and country-specific variables for 103 countries, and these countries are listed in Appendix 1. The overall index, which is calculated out of 600, of the *URAP* world ranking values was used for the rank variable. We assigned the rank variable by calculating the average score for each country, jointly entering in 2013–2019, in the *URAP* world rankings. The *political\_stability* variable represents political stability and absence of violence/terrorism and the *rule\_of\_law* variable reflects the degree of confidence in the rules of society and the quality of the courts. These two variables were used in percentile rank among all countries and obtained from the *Worldwide Governance Indicators (WGI) Project*. The *press\_freedom* variable is the global score of the world press freedom index ranging from 0 to 100 (the higher the score, the worse the freedom) and was retrieved from the *Reporters Without Borders (RSF)*. The *economic\_freedom* variable explains how much institutions correlate with the socioeconomic indicators. It was provided by the *Fraser Institute* and represents the economic freedom summary index, which is in the range 1–10 (best). The *industry\_col* variable expresses the university-industry collaboration on research and development in the range 1–7 (best) and was retrieved from the *World Economic Forum*. The *GDP* variable is the gross domestic product per capita in terms of constant US\$ for the base year 2010 and the data was obtained from the *World Bank*.



**Fig. 2** The countries used in the analysis

Because of the lack of data, it could not be evaluated for the effects of government expenditure on education, research and development expenditure, and human development index. Also, the countries Belarus, Cuba, Fiji, Iraq, Macao, Puerto Rico, Reunion, Sudan, and Taiwan could not be included in the analysis due to missing observations in the independent variables. The countries included in the analysis are illustrated in Fig. 2 in red color.

Figure 3 demonstrates the graphics belonging to the variables. It is seen that the variability of the *RANK* variable, which expresses the success ranking of the universities, is the highest compared to the other variables.

The analysis was firstly started with the Hausman test to determine its specific effects. As a result of the Hausman test, the null hypothesis could not be rejected. Therefore, it is concluded that random effects are valid in models. After determining that the models with random effects were appropriate in terms of specific effects, spatial models with random effects were estimated to perform tests on spatial coefficients. For this reason, the models of SAR, SDM, SEM, and GSPRE were estimated by random effects, respectively, and the estimation results are demonstrated in Table 1.

In the next stage, the model selection tests were carried out on the spatial coefficients to determine the most suitable model. The results of the tests are summarized in Table 2. Considering the  $\chi^2$  test statistics in model selection, first, the SAR and SDM models were compared; in consequence of this comparison, the null hypothesis was rejected and the SDM model was considered appropriate. After that, SEM and SDM models were compared and again it was decided that the SDM model was suitable. Finally, GSPRE and SDM models were compared considering the AIC criterion. Since the GSPRE model has a slightly smaller AIC value, it should be

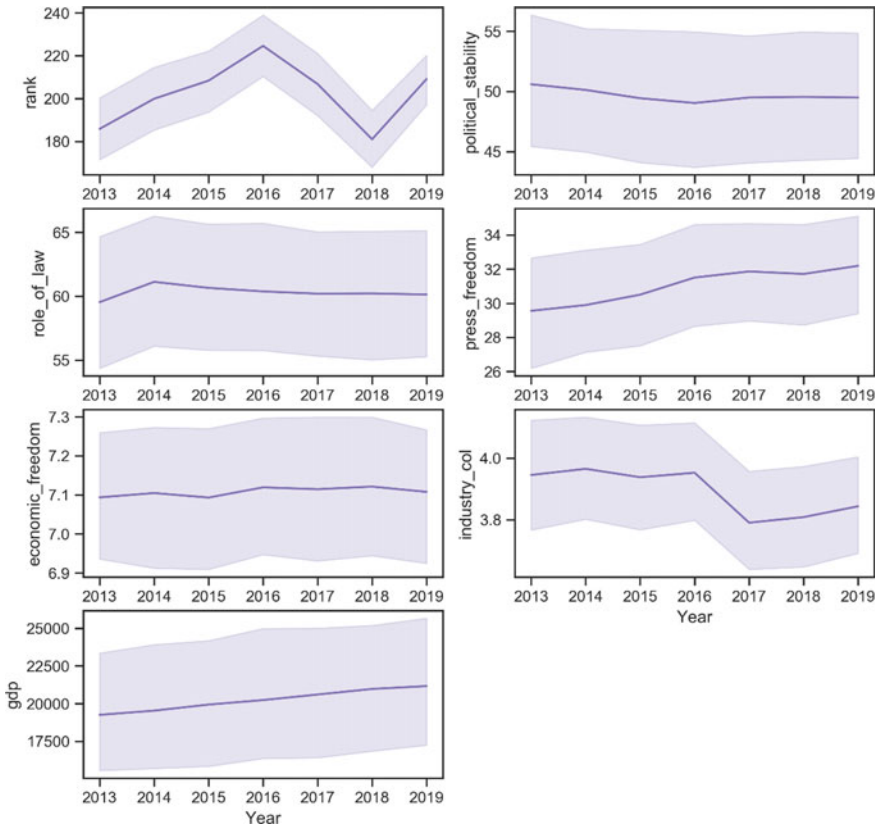


Fig. 3 The graphics of the variables

preferred as the suitable model. However, since no coefficients of the GSPRE model were statistically significant, the SDM with the random effects model was selected as the most suitable model.

When the findings of the SDM model are examined, it is concluded that the coefficients of the rule of law, the industry collaboration, the GDP per capita variables, and the constant are statistically significant. Also, it is seen that the spatial autocorrelation coefficient  $\rho$  is statistically significant in the model. When all findings are evaluated together, the conclusion can be drawn that the adoption of the rule of law principle in a country and being at the high level of the rule of law raises the rank of universities. In other words, it is possible to state that the rule of law is an important factor in terms of universities to be successful in the URAP ranking. The coefficient of university-industry collaboration index is also positive and statistically significant; a higher rate shows that this collaboration has an important effect on the university

**Table 1** Random effects model for the university ranking

	Random effects	SAR	SDM	SEM	GSPRE
political_stability	-0.389*	-0.366	-0.366	-0.349	-0.349
rule_of_law	0.809***	0.837***	0.741**	0.802**	0.804
press_freedom	-0.408	-0.365	-0.489	-0.563	-0.562
economic_freedom	-5.622	-5.165	-6.345	-6.522	-6.531
industry_col	13.444***	13.574***	14.479***	12.836***	12.836
Gdp	0.001***	0.001***	0.001***	0.001***	0.001
Constant	148.930***	130.605***	163.920***	161.915***	161.856
$\rho$		0.075**	0.148***		
$\lambda$				0.186***	0.186***
$\phi$					-0.005
Log-likelihood/Wald	101.26***	-3660.149	-3651.343	-3654.212	-3654.211
$R_w^2$	0.0089	0.0088	0.021	0.008	0.008
$R_b^2$	0.5008	0.4735	0.500	0.503	0.503
$R^2$	0.4229	0.4000	0.424	0.425	0.425
Hausman $\chi^2$	5.21	22.75	16.58	2.32	
Hausman p-value	0.5168	0.0019	0.219	0.940	

Note \*\*\*, \*\*, and \* denote rejections of the null hypothesis at the 1, 5 and 10% significance levels, respectively

**Table 2** Testing for model selection

Model	$\chi^2$	p-value	AIC
SAR vs SDM	19.21	0.0038	–
SEM vs SDM	13.25	0.0392	–
SDM	–	–	7334.685
GSPRE	–	–	7330.422

rankings. Finally, it is possible to state that the increase in per capita GDP value has an important place in terms of the success of universities. On the other hand, statistically significant findings on the variables of political stability, press freedom, and economic freedom in the model could not be reached.

## 4 Conclusion

The main motivation in our study is to reveal how the country-specific variables affect the URAP global rating scores. It analyzed data from 103 countries covering the period 2013–2019 by using the spatial panel models. Choosing the URAP index, which gives ranking information of 2500 universities, made a comprehensive study possible. It is thought that the study will make a significant contribution to the literature due to the examination of the relationship of the country-specific variables with a considerable number of countries and the evaluation of this relationship with spatial panel data methods. In the study, in which the most meaningful findings were obtained from the spatial Durbin model with random effects, it was concluded that the rule of law, university-industry collaboration, and per capita gross domestic product variables were positive and had significant effects on the URAP ranking.

The rule of law refers to the height of the spread and authority of law in a country. This concept, which expresses the importance of the rule of law especially against the state and those holding the power of government, also means that every citizen can be the addressee of the law. Therefore, the fact that the rule of law has a significant effect on the academic rankings of universities is quite suitable for expectations. In a country where the rule of law is not accepted, universities will not be expected to show success, as the improvement rate of science will be insufficient. When the university-industry collaboration variable providing another meaningful finding is examined, it should focus primarily on the benefits of universities' collaboration with industry. It is possible to state that university-industry collaboration benefits by improving the scientific research environment of universities, and increasing the quality of education and financial resources. Therefore, the high rate of this ratio will increase the quality of universities and will make a positive contribution to their success. Finally, considering the finding that the gross national product per capita variable has a significant effect on university success degrees, it is possible to state that per capita income is the criterion used by economists and governments to determine the general welfare level of a country, and thus the findings obtained are in line with economic expectations.

As a result, it can be emphasized that it is important for universities to cooperate with the industry to get higher scores in the URAP index. In terms of countries, it is possible to state that having a justice system that protects the rule of law and a high level of welfare provides success for universities. If both universities and countries consider the stated results, there is a possibility that their universities will achieve a more successful degree.



## Appendix

See Table 3.

**Table 3** The Countries Included in the Analysis

Algeria	Estonia	Luxembourg	Serbia
Argentina	Ethiopia	Malawi	Singapore
Armenia	Finland	Malaysia	Slovakia
Australia	France	Malta	Slovenia
Austria	Georgia	Mauritius	South Africa
Azerbaijan	Germany	Mexico	South Korea
Bahrain	Ghana	Montenegro	Spain
Bangladesh	Greece	Morocco	Sri Lanka
Belgium	Hong Kong	Namibia	Sweden
Bolivia	Hungary	Netherlands	Switzerland
Bosnia and Herzegovina	Iceland	New Zealand	Tanzania
Botswana	India	Nigeria	Thailand
Brazil	Indonesia	North Macedonia	Tunisia
Bulgaria	Iran	Norway	Turkey
Cameroon	Ireland	Oman	Uganda
Canada	Israel	Pakistan	Ukraine
Chile	Italy	Peru	United Arab Emirates
China	Jamaica	Philippines	United Kingdom
Colombia	Japan	Poland	United States
Costa Rica	Jordan	Portugal	Uruguay
Croatia	Kazakhstan	Qatar	Venezuela
Cyprus	Kenya	Romania	Viet Nam
Czech Republic	Kuwait	Russia	Yemen
Denmark	Latvia	Rwanda	Zambia
Ecuador	Lebanon	Saudi Arabia	Zimbabwe
Egypt	Lithuania	Senegal	

## References

- Alaşehir O (2010) University ranking by academic performance: A scientometrics study for ranking world universities. The middle east technical university.
- Alaşehir O, Çakır MP, Acartürk C, Baykal N, Akbulut U (2014) URAP-TR: a national ranking for Turkish universities based on academic performance. *Scientometrics* 101:159–178. <https://doi.org/10.1007/s11192-014-1333-4>
- Anselin L (2001) Spatial Econometrics. In: Baltagi BH (ed) *A companion to Theoretical Econometrics*. Blackwell Publishing Ltd., pp 310–330
- Anselin L, Gallo JL, Jayet H (2008) Spatial Panel Econometrics. In: Matyas L, Sevestre P (eds) *The Econometrics of Panel Data*. Springer-Verlag, Berlin, Heidelberg, pp 625–660
- Arimoto A (2011) Reaction to Academic Ranking: Knowledge Production, Faculty Productivity from an International Perspective. In J. C. Shin, R. K. Toutkoushian, & U. Teichler (eds), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education* (p. 271). Springer Science and Business Media. <http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf>
- Belotti F, Hughes G, Mortari AP (2017) Spatial Panel Data Models Using Stata. *Stand Genomic Sci* 17(1):139–180. <https://doi.org/10.2139/ssrn.2754703>
- Bi et al. (2018) Modeling spatiotemporal heterogeneity of customer preferences in engineering design. *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2A: 44th D*(August), 1–12. <https://doi.org/10.1115/DETC2018-86245>
- Boulton G (2011) University rankings: Diversity, excellence and the European initiative. *Procedia—Social and Behavioral Sciences*, 13, 74–82. <https://doi.org/10.1016/j.sbspro.2011.03.006>
- Çakır MP, Acartürk C, Alaşehir O, Çilingir C (2015) A comparative analysis of global and national university ranking systems. *Scientometrics* 103(3):813–848. <https://doi.org/10.1007/s11192-015-1586-6> Cliff, A.D., & Ord, K.J. (1973). *Spatial Autocorrelation*. Pion, London
- Cliff AD, Ord KJ (1973) *Spatial Autocorrelation*. Pion, London
- Elhorst JP (2011) *Spatial Panel Data Models*.
- Elken M, Hovdhaugen E, Stensaker B (2016) Global rankings in the Nordic region: challenging the identity of research-intensive universities? *High Educ* 72:781–795. <https://doi.org/10.1007/s10734-015-9975-6>
- Fraser Institute. (n.d.). Retrieved October 27, 2020, from <https://www.fraserinstitute.org/economic-freedom/dataset%3Fgeozone%3Dworld%26page%3Ddataset%26min-year%3D1970%26max-year%3D2018%26filter%3D0%26date-type%3Drange>
- Frenken K, Heimeriks GJ, Hoekman J (2017) What drives university research performance ? An analysis using the CWTS Leiden Ranking data. *J Informet* 11:859–872. <https://doi.org/10.1016/j.joi.2017.06.006>
- Geary RC (1954) The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician* 5(3):115–146
- Gerkman L (2010) Topics in Spatial Econometrics—With Applications to House Prices. In *Svenska handelshögskolan*. Hanken School of Economics.
- Goglio V (2016) One size fits all? A different perspective on university rankings. *J High Educ Policy Manag* 38(2):212–226. <https://doi.org/10.1080/1360080X.2016.1150553>
- Grewal R, Dearden JA, Lilien GL (2008) The university rankings game: Modeling the competition among universities for ranking. *Am Stat* 62(3):232–237. <https://doi.org/10.1198/000313008X332124>
- Günel FE (2018) Mekansal Panel Veri Modelleri. In S. Güriş (ed), *Uygulamalı Panel Veri Ekonometrisi* (1st ed., pp. 172–182). Der Yayınları, İstanbul.
- Günel Günel G, Altuğ G (2016) Avrupa Birliği Beşinci Genişleme Sürecinde İşsizlik Oranlarının Belirleyicileri: Mekansal Ekonometri Analizi. *Ege Stratejik Araştırmalar Dergisi* 7(2):237–252
- Hammarfelt B, de Rijcke S, Wouters P (2017) From Eminent Men to Excellent Universities: University Rankings as Calculative Devices. *Minerva* 55:391–411. <https://doi.org/10.1007/s11024-017-9329-x>

- Huang M-H (2012) Opening the black box of QS world university rankings. *Research Evaluation* 21:71–78. <https://doi.org/10.1093/reseval/rvr003>
- Ishikawa M (2009) University rankings, global models, and emerging hegemony: Critical analysis from Japan. *J Stud Int Educ* 13(2):159–173. <https://doi.org/10.1177/1028315308330853>
- Johnes J (2018) University rankings : What do they really show ? *Scientometrics* 115:585–606. <https://doi.org/10.1007/s11192-018-2666-1>
- Kehm BM (2014) Global university rankings—Impacts and unintended side effects. *Eur J Educ* 49(1):102–112. <https://doi.org/10.1111/ejed.12064>
- Kivinen O, Hedman J (2008) World-wide university rankings: A Scandinavian approach. *Scientometrics* 74(3):391–408. <https://doi.org/10.1007/s11192-007-1820-y>
- Kopczewska K, Kudła J, Walczyk K (2017) Strategy of Spatial Panel Estimation: Spatial Spillovers Between Taxation and Economic Growth. *Appl Spat Anal Policy* 10(1):77–102. <https://doi.org/10.1007/s12061-015-9170-2>
- Liu NC, Liu L (2005) University rankings in China. *High Educ Eur* 30(2):217–227. <https://doi.org/10.1080/03797720500260082>
- Marginson S (2007) Global university rankings: Implications in general and for Australia. *J High Educ Policy Manag* 29(2):131–142. <https://doi.org/10.1080/13600800701351660>
- Marginson S (2009) Open source knowledge and University rankings. In *Thesis Eleven* (Vol. 96, Issue 1). <https://doi.org/10.1177/0725513608099118>
- Marginson S (2014) University rankings and social science. *Eur J Educ* 49(1):45–59. <https://doi.org/10.1111/ejed.12061>
- McAleer M, Nakamura T, Watkins C (2019) Size, internationalization, and university rankings: Evaluating and predicting Times Higher Education (THE) data for Japan. *Sustainability* 11(1366):1–12. <https://doi.org/10.3390/su11051366>
- Meseguer-Martinez A, Ros-Galvez A, Rosa-Garcia A (2019) Linking YouTube and university rankings: Research performance as predictor of online video impact. *Telematics Inform* 43. <https://doi.org/10.1016/j.tele.2019.101264>
- Mingers J, O’Hanley JR, Okunola M (2017) Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics* 113:1627–1643. <https://doi.org/10.1007/s11192-017-2532-6>
- Moed HF (2017) A critical comparative analysis of five world university rankings. *Scientometrics* 110:967–990. <https://doi.org/10.1007/s11192-016-2212-y>
- Mori M (2016) How do the scores of world university rankings distribute? *5th IIAI International Congress on Advanced Applied Informatics*, 482–485. <https://doi.org/10.1109/IIAI-AAI.2016.36>
- Pietrucha J (2018) Country-specific determinants of world university rankings. *Scientometrics* 114:1129–1139. <https://doi.org/10.1007/s11192-017-2634-1>
- Pusser B, Marginson S (2013) University Rankings in Critical Perspective. *The Journal of Higher Education* 84(4):544–568. <https://doi.org/10.1080/00221546.2013.11777301>
- Rauhvargers A (2013). *Global University Rankings and Their Impact: Report II*. <https://doi.org/10.4324/9780203842171>
- Reporters Without Borders (RSF)*. (n.d.). Retrieved October 27, 2020, from <https://rsf.org/en/ranking/>
- Saka Y, Yaman S (2011) Üniversite Sıralama Sistemleri; Kriterler ve Yapılan Eleştiriler. *Journal of Higher Education and Science* 1(2):72–79. <https://doi.org/10.5961/jhes.2011.012>
- Salima BA, Julie LG, Lionel V (2018) Spatial econometrics on panel data. In V. Loonis & M.-P. de Bellefon (eds), *Handbook of Spatial Analysis: Theory and Application with R* (Issue 7, pp. 179–203).
- Sebetci Ö, Aksu G (2019) Comparison of academic structure of universities in Turkey by multiple correspondence analysis method. *International Journal of Learning, Teaching and Educational Research*, 18(6), 39–54. <https://doi.org/10.26803/ijlter.18.6.3>
- Teichler U (2011) Social contexts and systemic consequence of university rankings: A Meta-Analysis of the Ranking Literature. In J. C. Shin, R. K. Toutkoushian, & U. Teichler (eds),

- University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education (p. 271). Springer Science and Business Media.
- URAP. (n.d.). URAP Research Laboratory. Retrieved October 1, 2020, from <https://www.urapcenter.org/Methodology>
- US News and World Report. (n.d.). Retrieved January 16, 2021, from <https://www.usnews.com/info/articles/2018/06/11/celebrating-85-years>
- Usher A, Savino M (2007) A global survey of university ranking and league tables. High Educ Eur 32(1):5–15. <https://doi.org/10.1080/03797720701618831>
- Uslu B (2018) Dünya Üniversiteler Sıralaması : Genişletilen Gösterge Setine Göre Sıralamada Oluşan Farklılıklar the World University Rankings : Differentiations in Rankings According to the Expanded Indicator Set. 8(3):457–470. <https://doi.org/10.5961/jhes.2018.287>
- World Bank. (n.d.). Retrieved October 27, 2020, from <https://data.worldbank.org/>
- World Economic Forum. (n.d.). Retrieved October 27, 2020, from <http://reports.weforum.org/global-competitiveness-index-2017-2018/competitiveness-rankings/#series=EOSQ072>
- Worldwide Governance Indicators (WGI) project. (n.d.). Retrieved October 27, 2020, from <https://info.worldbank.org/governance/wgi/>
- Yonezawa A, Nakatsui I, Kobayashi T (2002) University Rankings in Japan. High Educ Eur 27(4):373–382. <https://doi.org/10.1080/0379772022000071850>

# The Impact of Outsourcing and Innovation on Industry 4.0



Ferda Esin Gülel  and Öncü Yanmaz Arpacı

**Abstract** In this study, we examined the impact of outsourcing and innovation activities of industry 4.0, focusing on the Turkish Fortune 500 list of companies. The list contains the largest companies in Turkey within the manufacturing, trade, and service industries. We obtained the data from the innovation-outsourcing scale adapted by Yanmaz Arpacı and Gülel (Yanmaz Arpacı Ö, Gülel FE (2019) Development of innovation-outsourcing scale in enterprises. In: Paper presented at 3rd international EUREFE congress. Adnan Menderes University, Aydın). This scale consists of four dimensions: the importance of outsourcing, the results of outsourcing, satisfaction in outsourcing, and innovation. We analyzed the data by using binary logistic and multinomial logistic regression methods. In the first model, the dependent variable is the status of having a research and development department in the company. In the second model, the dependent variable is taken as the company's state to switch to industry 4.0. As a result, we found that the importance of outsourcing and innovation is statistically significant in being research and development department. It must be noted throughout this research that the results of outsourcing, satisfaction in outsourcing, and innovation factors are statistically significant for industries switching to Industry 4.0.

**Keywords** Outsourcing · Innovation · Research and development · Industry 4.0

## 1 Introduction

Outsourcing is a strategic management approach that provides flexibility in production and capacity planning to companies. Outsourcing can make your business more flexible and agile, able to adapt to changing market conditions and challenges, while providing cost savings and service level improvements. Outsourcing provides attainment to new technologies, capacities, experiences, and knowledge (see p. 76–77 in Akgemci 2008). Today, companies can gain a competitive advantage when they

---

F. E. Gülel (✉) · Ö. Y. Arpacı  
Pamukkale University, Denizli, Turkey  
e-mail: [fegulel@pau.edu.tr](mailto:fegulel@pau.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_14](https://doi.org/10.1007/978-3-030-85254-2_14)

235

procure activities other than the core competencies from external sources specialized in these areas (see p. 184 in Inci and Acer 2019). Another critical factor in providing a competitive advantage in the economies of developed countries is innovation. The economist Schumpeter asks us to consider the importance of innovation by considering the impact innovation has as presented in his “business” growth model within his literature. Schumpeter highlights the importance of innovation that enables the development of processes in all “business” sectors (see p. 2 in Aghion et al. 2019). According to Schumpeter (see p. 32 in 1967), innovation is the commercial or industrial use of a new product, process or production technique, the emergence of an original market supply source, and the reshaping of organizations and companies.

Today, innovation arises as an approach and conception that generates potential solutions for institutions, individuals, sectors, and countries. The main factors affecting the success of innovation are the companies’ internal and external resources and the ability to manage them correctly. In this context, successful innovation is expected to cover both the company’s technical needs and sectoral needs. Innovation consists of many different dimensions and activities, such as a new product, a new production model, and cost reduction, redesigning production methods, and arranging distribution channels for a more efficient logistics activity. Companies carry out their innovation activities with actors in the supply chain in constantly growing sectors when considering outsourcing and innovation. The reason why innovation was rarely seen and progressed slowly in the past is that cooperation and supply chains were not as effective as today (see p. 174–200 in Kline and Rosenberg 2010).

In this study, we observed that companies practiced innovation in three ways. The first was to provide traditional outsourcing and to realize innovation within the company. The second was for the company to procure innovative outsources and recognize the innovation in its process, and the third to outsource the innovation directly. Our study’s subject is to examine the impact of outsourcing and innovation activities on industry 4.0 in the Turkish Fortune 500 list of companies. The research is essential in terms of determining the importance of outsourcing, the results of outsourcing, satisfaction in outsourcing, and innovation factors on the status of companies having a research and development (R&D) department and transition to Industry 4.0.

## 2 Literature Research

Yanmaz Arpacı (2019) identified the importance of outsourcing, the results of outsourcing, satisfaction in outsourcing, and innovation factors from her studies of Fortune 500 Turkish companies. In literature reviews, comprehensive research has not been found that discusses these factors, the existence of R&D departments of companies, and their transition to Industry 4.0. It is often the case that studies generally focus on the effect of each factor on R&D and Industry 4.0 in general. While the impact of the relevant factors on R&D and Industry 4.0 are discussed loosely in the

studies within the literature world, in this study, the effects of each of the factors on the company-based R&D activities and the transition to Industry 4.0 are examined.

Jones (see p. 341–351 in 2000) obtained data from two primary sources, these sources argued that technological innovations played a central role in the emergence of modern industrial societies. The first source of data used in Jones's study was from the UK R&D Scoreboard published annually by the Ministry of Trade and Industry, and the second was the commercial R&D initiatives research, which analyzed the appropriate initiatives annually by the National Statistical Office. In this study, technological developments and the internal or external status of R&D units were examined through the UK pharmaceutical industry sample. The data determined that R&D outsourcing has increased since the 1990s in the UK pharmaceutical industry. Large companies in the pharmaceutical industry prefer to invest in know-how from smaller companies and universities rather than investing in R&D directly.

Karplus (see p. 1–30 in 2007) defined in his case study focusing on the Chinese energy sector that innovation performance is higher within the corporate sector and enterprises. An institutional and robust R&D department is useful in companies' innovation activities to increase competition and transformation; this has been supported by clear evidence that a country's ability to be a productive partner in international collaborations is based on its R&D studies and initiatives' strength. Companies receive support from universities and research institutes to assist or substitute their in-house R&D. However, it has been observed that this external support is not always very efficient. It has been identified that creating resources for the development of R&D departments, especially in large-scale companies, is more coordinated and efficient than outsourcing R&D. The case study determines that both the general industry and R&D policy of the Chinese Government play an essential role in creating the conditions in which technological innovations can thrive and develop.

Grimpe and Kaiser (see p. 1483–1484 in 2010) proved that there is an inverse U-shaped relationship between outsourcing R&D and innovation performance in their study based on the data of innovative companies in Germany. There is a non-linear but positive relationship between the level of companies' internal R&D activities and the number of formal R&D collaborations and outsourcing. As the procurement of R&D activities has become a common practice, the question of whether external resources can positively affect innovation performance has gained importance. It has become a necessity for companies to balance the gains from R&D procurement and possible losses. As a result of these findings, it is necessary to analyze the expenditures for outsourcing in R&D with innovation expenditures in order to determine the main benefits of business management. It is also crucial for business management to know how to avoid the adverse effects of over-outsourcing and the critical threshold in an inverse U-shaped relationship.

In Harris and Moffat's study (see p. 1–19 in 2011), the relationship between R&D studies of companies and innovation was discussed through the UK Community Innovation Survey (CIS). The data was obtained from the results of the 2005, 2007, and 2009 surveys conducted by the British Office for National Statistics (ONS) and the innovation activity reports of the companies 2002–2008. In the study, companies from both the production and service sectors were examined. The study revealed that

46% of the manufacturing companies identified carried out product and/or process innovations through their R&D departments. In the service sector, the relationship between R&D and innovation is not as high as expected.

Fidanboy (2016) has asked to measure the impact of core competencies in the R&D performance of Turkish IT companies located in techno parks. Data was obtained from a total of 152 participants selected from 12 different companies operating in four different techno parks. The data was then evaluated by using a correlation and regression analysis model. In the context of the research results, there is a positive and significant relationship between companies' core competencies and R&D performance. The study is also intended to contribute to national R&D policies, R&D management, and governance.

Strange and Zucchella (see p. 174 in 2017) aimed to determine how new digital technologies such as the Internet of Things, big data, and robotic systems affect activities and organizations in global chains in the context of Industry 4.0. The approach taken in this research is to review various sources on the subject. In this context, Industry 4.0 was analyzed and compared with existing technologies, new technologies, and new configurations consisting of companies, customers, and suppliers being evaluated as a whole. The research concluded that Industry 4.0 is still in its infancy, but even in infancy it does have an impact on competition and corporate strategies in many sectors.

Vacik and Spacek (see p. 352 in 2018) surveyed mid-level managers in the Czech and Slovak Republic in their study. They concluded that companies are more interested in service innovation and developments devoted to industrial big data. The study's main output is a normative model design for growth related to outsourcing in service innovation. Both financial and non-financial criteria are examined within the model. Eighty-nine companies participated in the survey. The questionnaire was completed with contextual interviews with managers. The research has confirmed that companies are not sufficiently experienced to predict the future role of corporate outsourcing in the context of changes that appear with Industry 4.0.

Yıldız (see p. 1215–1223 in 2018) researched the connection and relationship of the digital supply chain concept and structure associated with Industry 4.0 in his literature review. The study also evaluates the importance of outsourcing and supply chains integrally within the scope of the internet of things and cloud computing. Digital supply chains, which replace traditional supply chains, have features such as reassigning alternative routes with logistical flexibility, controlling capacity with capacity flexibility, controlling the supply and delivery times, and having similar principles to Industry 4.0.

Alcácer and Cruz-Machado (see p. 915–916 in 2019) examined the appropriate conditions for a transition to Industry 4.0 in companies, the effectiveness of Industry 4.0 on production, and technological systems for Industry 4.0. As a result of the investigation, the Internet of Things, combined with many advanced systems, can offer a wide range of possibilities for innovation and optimization for companies. For example, the integration of Industry 4.0 of an SME in the supply chain, allows minimization of the risks associated with project development, product launch, open innovation, and sharing innovation. Product and service innovations that typically require



challenging and long research studies can be realized more easily with technologies such as virtual reality and simulation.

Weking et al. (see p. 1–12 in 2020) focused on the impact of Industry 4.0 on product and process innovations in the context of its technological aspects. Thirty-two case studies were examined and analyzed to measure the effects of Industry 4.0; the outcomes of the study became an innovative business model for companies to measure themselves against. Research is built on three main research designs: integration, service provision, and originality. Research findings have amplified current approaches to the role of Industry 4.0 on the ecosystem and how it affects business model innovations and service systems. The classified findings allow companies to evaluate their current business models in terms of availability during the transition to Industry 4.0.

### 3 Data

In this study, we analyzed the data obtained from the innovation-outsourcing scale adapted by Yanmaz Arpacı and Gülel (2019). Data was collected from 250 companies on Fortune 500 Turkey list. We examined the effects of the importance of outsourcing (O1), results of outsourcing (O2), satisfaction in outsourcing (O3), and innovation (I) dimensions on the status of being an R&D department in enterprises and switching to industry 4.0. We used the total scores of the answers given by the companies to these questions as data.

### 4 Method

In the analysis, we estimated using binary logistic and multinomial logistic regression methods. In cases where the dependent variable is categorical, and the independent variables are categorical or metric, the most appropriate model estimation method is logistic regression analysis (Hair et al. 2006). Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical dependent variable and one or more categorical or continuous independent variables.

When the dependent variable has two categories, the logistic regression model is shown in the figure below (Hosmer et al. 2013):

$$\ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 X + u$$

where  $u$  is an error term,  $\beta$  is a constant term representing unknown parameters. In the first model, while  $\pi(x)$  is a probability of having an R&D department in the company,  $1 - \pi(x)$  is a probability of not having an R&D department in the company.

When the dependent variable is multinomial (for example, four categories such as A, B, C, and D, D being the reference category), the model’s representation is as follows (Park and Kerr 1990):

$$\ln(\pi(A)/\pi(D)) = \beta_{AD} + \beta_{AD}X + u_{AD}$$

$$\ln(\pi(B)/\pi(D)) = \beta_{BD} + \beta_{BD}X + u_{BD}$$

$$\ln(\pi(C)/\pi(D)) = \beta_{CD} + \beta_{CD}X + u_{CD}$$

In the second model, the dependent variable is taken as the state of the company to switch to industry 4.0. The independent variables in both models are the total scores of the answers given to the importance of outsourcing (O1), outsourcing results (O2), satisfaction in outsourcing (O3), and innovation (I).

Maximum likelihood (Hosmer et al. 2013) can be used in binary logistic regression estimation, and generalized least squares (Theil 1970) or maximum likelihood (Nerlove and Press 1973) can be used in multinomial logistic regression estimation.

The categories of the dependent variables of the estimated model are given in Table 1.

In practice, researchers can see that logit and probit models are used for categorical dependent variables. The most important difference between probit and logit models are in the distribution of the model’s error term. While the error term is normally distributed in the probit model, it has a logistic distribution in the logit model. In addition, the calculation and the interpretation of the predicted probit model are more complex than the logit model (see p. 117–136 in Azen and Walker 2011). Therefore, we have used logit model in our research due to its assumption, calculation, and interpretation advantages.

**Table 1** Dependent variables in the model

	Categories	Frequency
Does your company have an R&D department?	Yes	156
	No (Reference)	94
What is the current status of your company regarding Industry 4.0?	We completely switched to Industry 4.0	71
	We have partially switched to Industry 4.0	84
	We are at the planning stage regarding Industry 4.0	25
	We do not intend to switch to Industry 4.0 (Reference)	66

## 5 Findings

Descriptive statistics of the independent variables in the analysis according to the dependent variable categories are given in Table 2.

In the first stage of our research, we determined the dependent variable as 0 if there is not an R&D department in the company, and 1 if there is. We have given the model estimates we obtained according to the backward method in Tables 3, 4, 5, and 6.

The likelihood ratio Chi-square test indicates that our full model is a significant improvement in fit over a null (intercept-only) model. The  $\chi^2$  value of the model was calculated as 21.918. The model became statistically significant in the third stage according to the backward method ( $p < 0.05$ ).

The Hosmer and Lemeshow test indicates goodness-of-fit. The test value Chi-square is 3.858 ( $p > 0.05$ ). It shows that the model adequately fits the data.

In the classification table, the percentage correct associated with row No is referred to as specificity, as it reflects the accuracy of the model in correctly classifying companies into group 0. The No-group indicates that the company does not have an R&D department.

The percentage correct associated with row Yes is referred to as sensitivity, as it reflects the accuracy of the model in correctly classifying companies into group 1. The Yes-group indicates that the company has an R&D department.

The correctly classifying percentage is 26.2% for No-group, as it is 87.2% for Yes-group. The overall percentage is 64.4%. The percentages are good, both Yes-group and overall. However, it is clear that the model might work better for the No-group.

The Wald tests in the tables show the significance of the coefficients of the model. The importance of outsourcing (O1) is negative and significant ( $\beta = -0.085$ ,  $S.E. = 0.040$ ,  $p = 0.032$ ). It is a predictor of the likelihood of a company having an R&D department. The odds ratio indicates that for every one-unit increase on the importance of outsourcing, the odds of having an R&D department decreases by 0.849. When this finding and the responses given by companies are examined, it shows that the status of having an internal R&D department decreases as outsourcing gains importance in companies. When companies prefer to outsource, and successful cooperation is established in the long term, they may prefer to procure R&D activities concurrently. It is thought that this preference will provide benefits to the company in terms of cost and efficiency.

Innovation is positive and significant ( $\beta = 0.087$ ,  $S.E. = 0.020$ ,  $p = 0.000$ ). It is a predictor of the likelihood of a company's innovation. The odds ratio indicates that for every one unit increase on the innovation, the odds of having an R&D department increases by 1.049. The finding obtained in line with the opinions of companies on innovation explains that as the innovation activities increase, the importance of R&D departments also increases. R&D is one of the essential foundations of innovation and can be more successful when managed institutionally and professionally. In this case, an internal R&D system can be considered to be more beneficial for companies.

**Table 2** Descriptive statistics of independent variables

		N	Min	Max	Mean	Std. deviation	
Does your company have an R&D department?	No	O1	8	35	26.39	4.192	
		O2	9	35	25.05	5.188	
		O3	8	29	21.86	4.089	
		I	94	19	59	43.16	8.794
	Yes	O1	156	14	35	26.22	3.739
		O2	156	7	35	25.56	4.936
		O3	156	6	30	21.89	4.533
What is the current status of your company regarding Industry 4.0?		I	156	12	60	47.49	7.209
	We completely switched to Industry 4.0	O1	8	33	25.86	4.196	
		O2	71	7	34	24.73	4.852
		O3	71	10	29	22.14	3.844
		I	71	12	59	45.27	8.159
	We have partially switched to Industry 4.0	O1	84	14	35	26.74	3.796
		O2	84	9	35	25.99	4.617
We are at the planning stage regarding Industry 4.0		O3	84	6	29	21.31	4.631
		I	84	19	59	45.18	8.292
		O1	25	20	35	25.72	3.298
		O2	25	9	31	24.00	5.545
		O3	25	14	29	22.48	4.043
		I	25	25	55	43.08	7.659
	We do not intend to switch to Industry 4.0	O1	66	9	35	26.11	3.828

(continued)

**Table 2** (continued)

		N	Min	Max	Mean	Std. deviation
	O2	66	12	35	25.21	5.293
	O3	66	8	30	22.06	4.271
	I	66	28	60	47.91	7.325

**Table 3** Omnibus tests of model coefficients

		Chi-square	df	Sig.
	Step	21.918	1	0.692
Step 3 <sup>a</sup>	Block	21.918	2	0.000
	Model	21.918	2	0.000

**Table 4** Hosmer and Lemeshow test

Step	Chi-square	df	Sig.
3	3.858	8	0.870

**Table 5** Classification table<sup>a</sup>

Predicted					
			Does your company have an R&D department?		
Observed			No	Yes	Percentage correct
Step 3	Does your company have an R&D department?	No	25	69	26.6
		Yes	20	136	87.2
Overall percentage					64.4

<sup>a</sup>The cut value is 0.500

**Table 6** Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 3 <sup>a</sup>	O1	-0.085	0.040	4.594	1	0.032	0.918	0.849	0.993
	I	0.087	0.020	18.854	1	0.000	1.091	1.049	1.135
	Constant	-1.206	1.054	1.309	1	0.253	0.299		

<sup>a</sup>Variable(s) entered on step 1: O1, O2, O3, I

Outsourcing of R&D carries the risk of sharing some crucial company-specific information with external sources. It is more advantageous for the company that an R&D department exists in the company in terms of defining innovations and providing a competitive advantage in the market. In addition, the companies that successfully innovate and gain significant earnings also increase their internal R&D investments in this context.

In the second phase of our research, we determined the dependent variable as 1 if the company “completely switched” to the industry 4.0, 2 if it “partially switched”, 3 if it was in “the planning stage”, and 4 if it was “do not intend to switch”. We

have given the model estimates we obtained according to the Stepwise method in the tables below:

Table 7 contains a Likelihood Ratio Chi-square test, comparing the full model (i.e., including all the predictors) against a null (or intercept only model). Statistical significance indicates that the full model represents a significant improvement in fit over a null model. In the table above, we see that the final model is significant ( $\chi^2 = 24.925, p < 0.05$ ).

Table 8 contains the Deviance and Pearson Chi-square test, which determine whether a model exhibits a good fit to the data. Non-significant test results are indicators that the model fits the data well. Both Deviance and Pearson’s Chi-square tests indicate adequate fit.

Table 9 contains likelihood ratio tests of the overall contribution of each independent variable to the model. We see that all independent variables are significant ( $p < 0.05$ ).

The results in Table 10 provide comparative information within each group against the reference category (we do not intend to switch to Industry 4.0). Specifically, the regression coefficients indicate which predictors significantly discriminate between

**Table 7** Model fitting information

	Model fitting criteria	Likelihood ratio tests				
Model	AIC	BIC	−2 Log likelihood	Chi-square	df	Sig.
Intercept only	648.195	658.711	642.195			
Final	641.269	683.333	617.269	24.925	9	0.003

**Table 8** Goodness-of-fit

	Chi-square	df	Sig.
Pearson	738.634	702	0.164
Deviance	614.497	702	0.992

**Table 9** Likelihood ratio tests

	Model fitting criteria	Likelihood ratio tests				
	AIC of reduced model	BIC of reduced Model	−2 Log likelihood of reduced model	Chi-square	df	Sig.
Intercept	638.529	670.077	620.529	3.260	3	0.353
O2	648.478	680.026	630.478	13.209	3	0.004
O3	646.838	678.386	628.838	11.568	3	0.009
I	644.781	676.329	626.781	9.512	3	0.023

**Table 10** Parameter estimates

									95% Confidence interval for Exp(B)	
		B	Std. error	Wald	df	Sig.	Exp(B)	Lower bound	Upper bound	
The current status of your company regarding Industry 4.0	Intercept	1.895	1.334	2.017	1	0.156				
	O2	-0.011	0.044	0.069	1	0.793	0.989	0.907	1.077	
	O3	0.029	0.052	0.316	1	0.574	1.029	0.93	1.139	
	I	-0.047	0.024	3.72	1	0.054	0.954	0.91	1.001	
We have partially switched to Industry 4.0	Intercept	2.184	1.311	2.776	1	0.096				
	O2	0.122	0.047	6.607	1	0.01	1.13	1.029	1.24	
	O3	-0.1	0.05	3.974	1	0.046	0.905	0.821	0.998	
	I	-0.062	0.024	6.652	1	0.01	0.94	0.896	0.985	
We are at the planning stage regarding Industry 4.0	Intercept	1.645	1.69	0.947	1	0.33				
	O2	-0.049	0.059	0.703	1	0.402	0.952	0.849	1.068	
	O3	0.099	0.074	1.796	1	0.18	1.104	0.955	1.275	
	I	-0.079	0.032	5.974	1	0.015	0.924	0.867	0.984	

<sup>a</sup>The reference category is: We do not intend to switch to Industry 4.0



“completely switched” and “not intend to switch”, between “partially switched” and “not intend to switch”, and between “the planning stage” and “not intend to switch”.

The first set of coefficients represent comparisons between “completely switched” and “not intend to switch”. O2 and O3 are not statistically significant ( $p > 0.05$ ), I is “near significant” ( $\beta = -0.047$ ,  $S.E. = 0.024$ ,  $p = 0.054$ ), as companies scoring higher on this variable are less likely to switch to Industry 4.0. The odds ratio of 0.954 indicates that for every one unit increase in innovation, the odds of company switching to Industry 4.0 decreases by 0.954. When the company responses given on the subject are examined, it can be thought that those companies who are successful in innovation do not want to switch to Industry 4.0 because they find themselves successful in this field. Some companies perceive Industry 4.0 as a disruptive innovation. For this reason, they are cautious about the transition to this system. Besides, companies that have achieved success in innovation think that they have a competitive advantage due to these successes and do not want to introduce a radical change. Another point of view is that Industry 4.0 is considered as an innovation itself. Companies that are currently innovating in other ways may not want to execute a new and large-scale innovation plan such as Industry 4.0 or may delay it to a later time.

The second set of coefficients represent comparisons between “partially switched” and “not intend to switch”. O2 ( $\beta = 0.122$ ,  $S.E. = 0.047$ ,  $p = 0.010$ ), O3 ( $\beta = -0.100$ ,  $S.E. = 0.050$ ,  $p = 0.046$ ) and I ( $\beta = -0.062$ ,  $S.E. = 0.024$ ,  $p = 0.010$ ) are statistically significant ( $p < 0.05$ ). O3 and I indicate that company’s score higher on these variables are less likely to partially switch to Industry 4.0. The odds ratio of 0.905 and 0.940 indicates that for every one unit increase on O3 and innovation, the odds of the company partially switching to Industry 4.0 decreases by 0.905 and 0.940. O2 indicates that company’s score higher on this variable are more likely to partially switch to Industry 4.0. The odds ratio of 0.010 indicates that for every one unit increase on O2, the odds of the company partially switching to Industry 4.0 increases by 0. Due to major industrial developments in the production sectors, companies included themselves in these rapid changes to compete and develop some strategies in this context. When considered integrally, Industry 4.0 is the main framework of all these strategies. In the context of this point of view, the increase in O3 factor (satisfaction in outsourcing) will enable the company to focus primarily on its core competencies. It is expected within strategic management philosophy that companies who focus on their core competencies will be more successful. Just like the innovation factor, businesses that see this situation as a competitive advantage do not want to take the risk of a radical change.

The third set of coefficients represent comparisons between “the planning stage” and “not intend to switch”. O2 and O3 are not statistically significant ( $p > 0.05$ ), I is a significant predictor ( $\beta = -0.079$ ,  $S.E. = 0.032$ ,  $p = 0.015$ ) in the model, as companies scoring higher on this variable are less likely to switch to Industry 4.0. The odds ratio of 0.940 indicates that for every one unit increase on innovation, the odds of the company planning stage to Industry 4.0 decreases by 0.940. The innovation dimension’s effect on Industry 4.0 has been discussed and evaluated in the first set of coefficients paragraph above in line with the results of Table 10.

**Table 11** Classification

Predicted					
Observed	We completely switched to Industry 4.0	We have partially switched to Industry 4.0	We are at the planning stage regarding Industry 4.0	We do not intend to switch to Industry 4.0	Percent correct (%)
We completely switched to Industry 4.0	21	37	0	13	29.6
We have partially switched to Industry 4.0	12	55	0	17	65.5
We are at the planning stage regarding Industry 4.0	10	11	0	4	0.0
We do not intend to switch to Industry 4.0	14	34	0	18	27.3
Overall percentage (%)	23.2	55.7	0.0	21.1	38.2

Table 11 is classification statistics used to determine which group memberships are best predicted by the model. “Completely switched” were correctly predicted by the model 29.6%. “Partially switched” were correctly predicted by the model 65.5%. “Not intend to switch” were correctly predicted by the model 27.3%. The model fails to correctly predict “the planning stage”. In other words, it is not a good predictor for “the planning stage”. Overall, 38.2% of the companies are classified correctly.

## 6 Results

In this study, we aim to measure the impact of outsourcing and innovation on Industry 4.0 in Fortune 500 companies in Turkey. Data was collected from 250 companies on Fortune 500 Turkey list. It has been observed that 60.9% of the companies studied have internal R&D departments. The research results indicate that 29.7% of the companies have completely or partially switched to Industry 4.0, and 28.9% are in the planning stage. We examined the effects of the importance of outsourcing (O1), results of outsourcing (O2), satisfaction in outsourcing (O3), and innovation (I) dimensions on the status of being an R&D department in enterprises and switching

to industry 4.0. We used the total scores of the answers given by the companies to these questions as data.

In the analysis, we estimated using binary logistic and multinomial logistic regression methods. The independent variables in both models are the total scores of the answers given to the importance of outsourcing (O1), outsourcing results (O2), satisfaction in outsourcing (O3), and innovation (I). In the first model, the dependent variable is the status of having an R&D department in the company. In the second model, the dependent variable is taken as the state of the company to switch to Industry 4.0.

Our models' results argue that the odds ratio indicates that for every one unit increase on the importance of outsourcing (O1), the odds of having an R&D department decreases by 0.849. This suggests that as outsourcing gains importance in companies, the situation of having an R&D department within the company was reduced. Companies prefer to outsource R&D as well as in other activities. Another result of the study, the results of outsourcing (O2) and the satisfaction in outsourcing (O3) are not statistically significant, innovation (I) is "near significant", as companies scoring higher on this variable are less likely to switch to Industry 4.0. The odds ratio of 0.954 indicates that for every one unit increase in innovation, the odds of company switching to Industry 4.0 decreases by 0.954. Although the findings display that innovative companies do not want to switch to Industry 4.0, it is thought that the transition to Industry 4.0, which is seen as a disruptive innovation, may be more comfortable, especially in companies that are already successful in their innovations. It is predicted that the transition to Industry 4.0, a new business model innovation, will be an irreplaceable condition for strong and competitive positions and sustainability, especially in the upcoming years.

Industry 4.0 is the main framework of the strategies developed exclusively concerning sectoral and technological developments in production. In case, outsourcing is a strategic management method. As satisfaction in outsourcing increases, companies can focus more on their core competencies. This provides many advantages to companies, in particular efficiency and productivity. As in the innovation dimension, companies that want to conserve their current competitive advantage can perceive a major and radical change as a risk at the transition to Industry 4.0.

The models composed in this study represent Turkey's largest companies' maturity levels in relation to strategic management and innovativeness in the context of the fourth industrial revolution. The approach agreed to carry out this study within both the production and service sectors was vital in terms of approaching the cases aggregately. We predict that the results obtained within the scope of both models will provide a perspective and contribution to outsourcing, innovation, R&D structuring, and Industry 4.0 project studies with different combinations. In future studies, it is recommended to perform the analysis by grouping the companies primarily as production-service operations and/or by the sectoral basis they are involved in.

## References

- Aghion P, Akcigit U, Bergeaud A, Blundell R, Hémous D (2019) Innovation and top income inequality. *Rev Econ Stud* 86(1):1–45
- Akgemci T (2008) *Stratejik Yönetim*. Gazi Kitabevi, Ankara
- Alcácer V, Cruz-Machado V (2019) Scanning the industry 4.0: a literature review on technologies for manufacturing systems. *Eng Sci Technol Int J* 22(3):899–919
- Azen R, Walker CM (2011) *Categorical data analysis for the behavioral and social sciences*. Routledge, London
- Fidanboy CÖ (2016) *Ulusal Ar-Ge politikaları bağlamında temel yetenek tabanlı Ar-Ge yönetimi yaklaşımı: teknokentler örneği*. Dissertation, Başkent University
- Grimpe C, Kaiser U (2010) Balancing internal and external knowledge acquisition: the gains and pains from R&D outsourcing. *J Manag Stud* 47(8):1483–1509
- Hair JF et al (2006) *Multivariate data analysis*. Pearson Prentice Hall, Upper Saddle River, NJ
- Harris R, Moffat J (2011) R&D, innovation and exporting. Available in Spatial Economics Research Centre (SERC). <http://eprints.lse.ac.uk/33593/1/sercdp0073.pdf>. Accessed 15 Nov 2019
- Hosmer JR, Lemeshow S, Sturdivant RX (2013) *Applied logistic regression*. Wiley, New Jersey
- Inci H, Acer A (2019) Lojistik faaliyetlerde dış kaynak kullanımı: karadeniz Bölgesi fındık işleticileri ve ihracatçıları üzerine bir uygulama. *Beykoz Akademi Dergisi* 7(2):183–201
- Jones O (2000) Innovation management as a post-modern phenomenon: the outsourcing of pharmaceutical R&D. *Br J Manag* 11(4):341–356
- Karplus VJ (2007) *Innovation in China's energy sector*. Center for Environmental Science and Policy, Stanford
- Kline SJ, Rosenberg N (2010) An overview of innovation. *Studies on science and the innovation process: selected works of Nathan Rosenberg*. World Scientific Publishing Co Pte Ltd., Singapore, pp 173–203
- Nerlove M, Press SJ (1973) *Univariate and multivariate log-linear and logistic models*, vol 1306. Rand, Santa Monica
- Park KH, Kerr PM (1990) Determinants of academic performance: a multinomial logit approach. *J Econ Educ* 21(2):101–111
- Schumpeter JA (1967) *Economic doctrine and method an historical sketch*. Galaxy Books, New York
- Strange R, Zucchella A (2017) Industry 4.0, global value chains and international business. *Multinatl Bus Rev*. <https://doi.org/10.1108/MBR-05-2017-0028>
- Theil H (1970) On the estimation of relationships involving qualitative variables. *Am J Sociol* 76(1):103–154
- Vacik E, Spacek M (2018) Process innovation in service sector matching industry 4.0 environment, In: Katalinic B (ed) *Proceedings of the 29th DAAAM international symposium*, Vienna, Austria, pp 0352–0360. <https://doi.org/10.2507/29th.daaam.proceedings.051>
- Weking J, Stöcker M, Kowalkiewicz M, Böhm M, Krcmar H (2020) Leveraging industry 4.0—a business model pattern framework. *Int J Prod Econ*. <https://doi.org/10.1016/j.ijpe.2019.107588>
- Yanmaz Arpacı Ö, Gülel FE (2019). Development of innovation-outsourcing scale in enterprises. In: Paper presented at 3rd international EUREFE congress. Adnan Menderes University, Aydın. <http://www.eurefe.com>. Accessed 1–3 Nov 2019
- Yanmaz Arpacı Ö (2019) *İşletmelerde dış kaynak kullanımı ve inovasyon ilişkisi*. Dissertation, Pamukkale University
- Yıldız A (2018) Endüstri 4.0 ile bütünleştirilmiş dijital tedarik zinciri. *Bus Manag Stud Int J* 6(4):1215–1230

# Subjective Well-Being of Poor Households



Süreyya Temelli  and Mustafa Sevüktekin

**Abstract** In the neoliberalizing world, social policy practices are declining. However, social assistance, which is one of the important tools of social policy, is crucial in terms of reducing poverty while also ensuring the reproduction of labor. There are limited number of studies investigating the influence of social assistance which helps poor people to meet their needs in-kind or in-cash on subjective well-being. Using 2013 Income and Living Conditions Survey from TURKSTAT, this study contributes empirically in this inquiry by looking at effects of the social assistance on subjective well-being. For this purpose, partial proportional odds model was used. According to the results, being recipient of social assistance has been found statistically significant as predictors of subjective well-being. Also, social assistance has a negative effect on subjective well-being. This outcome of the study suggests that people who receive social assistance feel poorer, therefore they report themselves less likely to be happy.

**Keywords** Subjective well-being · Social assistance · Partial proportional odds model

## 1 Introduction

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy-Anna Karenina

Since ancient Greece, the quality of life or the conditions for happiness has been widely discussed to improve welfare of community. While there are more than one definitions of well-being in the literature that make an individual happy, there is no consensus among these definitions. Consequently, measurement of well-being

---

S. Temelli (✉)

Department of Econometrics, Trakya University, Edirne, Turkey

e-mail: [sureyyadal@trakya.edu.tr](mailto:sureyyadal@trakya.edu.tr)

M. Sevüktekin

Department of Econometrics, Uludağ University, Bursa, Turkey

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

251

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_15](https://doi.org/10.1007/978-3-030-85254-2_15)

changes according to different opinions. As a result, different terms are used to define subjective well-being. These include quality of life, happiness and satisfaction. There are vast of studies on subjective well-being in order to understand needs for a good life. While the majority of the studies in the literature aim to determine the factors determining subjective well-being, there are a limited number of studies investigating the effect of social assistance on subjective well-being. Social assistance provides social security to individuals who are lack of contribution to welfare system. In this context, individual social assistance is provided in the form of monetary, in-kind or service, while the financing is provided by the state.

Killburn et al. (2018) analyzed the short-term impact of external positive income shock on caregivers' subjective well-being using a 17 month follow-up panel of 3365 households included in the Malawi outreach cash transfer program for extremely poor and labor-limited households. According to the results of the analysis, caregivers living in households receiving social assistance reported higher life satisfaction and tend to look more confidently into the future. However, the impact of social assistance on individuals' welfare in Turkey was investigated qualitatively by social scientists and the results of the studies have drawn attention to the problems on the welfare system. Güneş (2012) identified problem areas that prevented the continuation of social assistance. These are sustainability of social assistance, delivery of social assistance, increasing demand for social assistance, dependence of social assistance in the poor and social assistance in the form of political patronage. Kutlu (2015) investigated social assistance in Turkey in terms of the nature of social rights. His findings are based on field research study. In the study, it was revealed that beneficiaries of social welfare have lack knowledge about social assistance and could be involved in social assistance processes by their political relations. Kutlu also is a point to other problems which are the way of distribution of social assistance, the quality of the food assistance, uncertainties in the sustainability of social assistance and realization of the demand for social assistance in the form of social injustice. Taşçı (2017) mentioned three main concerns which were danger of stigmatization and humiliation of social assistance beneficiaries, addiction to laziness and exploit of social assistance. Karadoğan (2018) emphasized that social assistance contributes to poverty sustainability and that poverty is internalized through social assistance. According to the results of the study, social assistance which is not perceived as a social right affects the self-esteem level of individuals.

The above mentioned studies revealed the necessity to investigate quantitatively the effect of social assistance on individuals' quality of life in Turkey. For this purpose, this study, which is an effort to understand how the poor live in poverty in the transforming world, has taken a holistic approach by making use of the quantitative data to reach scientific conclusions and discussing results by utilizing the shared data of the qualitative studies mentioned above. In this way, not only the quantitative data was used to describe the society, but also the consistency of the information obtained through quantitative data was compared with qualitative observations. Thus, instead of preferring one method to another, the advantages of the two were utilized. In this context, the results obtained based on micro data can be used to determine policy at macro level. In order to achieve this, partial proportional odds model was used in

order to take into account ordered structure of dependent variable. Using this model is an attempt to make a contribution to use of ordinal utility theory to the possibility and the relevance of interpersonal comparisons for policy recommendations on welfare system.

The study is organized as follows. Section 2 presents the social assistance system in Turkey. Section 3 discusses the methodology. Section 4 summarizes data and presents the empirical results of the model which shows effect of social assistance on subjective well-being. Last section concludes.

## 2 Social Assistance System in Turkey

The social security system is a security system that provides protection against the physiological risks (illness, disability, old age, maternity, accidents and death) and economic risks (insufficient family income, unemployment) that individuals may face in life. These physiological and economic risks can lead to decrease in income, income cut or increases in expenses. Owing to the social security system, a minimum guarantee is provided for the person to fight against these risks (İzgi 2008). This guarantee is provided based on the paid premiums (social insurance) or non-contributory payments financed by taxes (public social security expenditure) (Akar 2015). The social assistance system is provided by non-contributory payments. The primary aim of social assistance system is to ensure social security to individuals who are lack of premium payment power. In this context, individual social assistance is provided in the form of monetary, in-kind or service, while the financing is provided by the state. Social assistance is distributed by Social Assistance and Solidarity Foundation with Law No. 3294 on the Encouragement of Social Assistance and Solidarity in Turkey since 1986 (Alper 2017).

Social assistance in Turkey can be divided into six main groups which are family assistance, shelter and food assistance, education assistance, health assistance, assistance for disabled person and old people, employment assistance and assistance for special purposes. The primary objective of these social assistance categories is to alleviate poverty by encouraging social justice. However, each of them achieves this goal using different tools. Family assistance provides a wide range of assistance from food needs to accommodation needs. In addition, this aid covers different types of households. These are military families in need, parentless children and widowed women. Although family assistance covers the need for shelter and food for those in need, shelter-food assistance to the benefit of households living in Turkey are also available. Shelter-food assistance provides electricity consumption support to households in need benefiting from social assistance programs. Also under social cohesion assistance which is part of shelter-food assistance, monthly payments are made to foreigners with temporary protection status, international protection status, international protection applicants and humanitarian residence permits. Education assistance contains free textbook distribution to the primary and secondary school students; free lunch for poor students who move to centers where schools are

located within the mobile education; food and accommodation help for primary and secondary school students outside the mobile education system; conditional education aid to families who do not have social security and who are in need within the scope of Law No. 3294 for their children's formal education on the condition that children do not have more than four days of absence in a month; dormitory construction for secondary school students and also transportation aid for students who need special education. Health assistance is provided to help people in need to meet their health needs. Some of the health assistance are premium payments to individuals without social security; providing all kinds of tools and equipment for disabled citizens in need; electricity consumption support for patients who are dependent on the device due to chronic illness; health assistance provided that families in need on the condition that send their children to health control; conditional health assistance for pregnant women in need provided that they have health checks and births in hospital; regular cash benefits for patients who experience psycho-social and financial loss due to tuberculosis and subacute sclerosing panencephalitis disease. In addition to covering the needs of the disabled person and old people within the scope of health assistance, cash assistance is also provided within the scope of assistance for them. As part of this assistance, monthly payments are also made to relatives of disabled people under the age of 18. In addition to these assistance programs, orientation and start-up assistance to a job for individuals who are able to work between the ages of 18–55 living in households benefiting from social assistance programs are provided within the scope of employment assistance. Apart from these aids, there are assistance programs for special purposes such as soup kitchens in poor neighborhoods, disaster and emergency aids, terrorist damage aids. According to the latest activity reports of the Republic of Turkey Ministry of Family, Labor and Social Services, regular social assistance in figures is shown in Table 1.

Regular social assistance programs excluding general health insurance premium support are conditional education assistance, conditional health assistance, conditional pregnancy assistance, assistance for widowed women, assistance for military families in need, assistance under Law No. 2022 and home care assistance. However, temporary assistance programs which are food assistance, fuel assistance, accommodation assistance, education assistance, health assistance, disability needs assistance, special purpose assistances, clothing and other family assistance, employment assistance are in the form of one-time assistance. As seen in Table 1, compared to 2014, the share of social assistance in the gross domestic product increased by 0.02% in 2016. The number of households benefiting from regular assistance programs in 2016 increased by 68,764 compared to 2014, while the number of staff responsible working in Social Assistance and Solidarity Foundation for the distribution of these aids increased by 447 and the number of Social Assistance and Inspection staff working in Social Assistance and Solidarity Foundation increased by 47. In addition, in 2016, the rate of individuals whose daily expenditure per capita was below 4.3 USD in purchasing power parity decreased by 0.48% compared to 2014. Therefore, in this study, housing aid, in-cash and in-kind social assistance, cash and in-kind child assistance were taken as social assistance and the effects of these on the subjective well-being were investigated using partial proportional odds model.



**Table 1** Overview of social assistance in Turkey

	2014	2015	2016
Total social assistance expenditure (thousand \$)	19,651,707.27	19,253,378.4	20,379,970.07
Share of social assistance expenditures in GDP*	1.06%	1.03%	1.08%
Number of households receiving social assistance	3,005,898	3,017,969	3,154,069
Number of households receiving regular social assistance	2,274,182	2,318,042	2,342,946
Number of households receiving temporary social assistance	1,892,656	1,924,649	2,046,888
Amount transferred to assistance from social assistance and solidarity encouragement fund (SYDTF) resources (\$)	3,956,182,672	3,882,557,212	3,662,837,142
Number of old age and disability salary beneficiaries under law no. 2022	1,300,377	1,272,038	1,292,355
Number of people for whom universal health insurance (UHI) contributions are paid by the government	9,368,920	8,983,853	6,683,106
Rate of individuals with per capita daily expenditure below 2.15 used per current purchasing power parity (PPP) (2013)	%0.06	%0.03	%0.06
Number of social assistance and solidarity foundations (SASF)	1000	1000	1000
Number of SASF staff	8611	8948	9058
Number of SASF social assistance and inspection officers	3792	3923	3839

\*Previous year values are taken for purchasing power parity

### 3 Empirical Methodology

Models for ordinal outcomes differ according to whether the distance between the categories is equal or not. Therefore, firstly, in the study, the parallel lines assumption was tested in order to test the equality of the distance between the categories of the outcome variable. Brant test was used for this purpose. With respect to the results of Brant test, some variables violate the parallel lines assumption. Therefore, the partial proportional odds model was used due to the ordered nature of the outcome variable.

Brant test is a Wald test that includes individual tests that show by which variables the assumption of parallel lines is violated. For this reason, firstly, j-1 binary logit was created for the outcome variable having J-category. The parallel lines hypothesis was tested as a result of the comparison of these binary logits (Long and Freese 2014). In the study, Eq. (1) was used to test parallel lines assumption for the outcome variable

with 11 category.

$$\Pr(\text{SWB} \leq j | \mathbf{x}) = F(\alpha_j - \mathbf{x}\beta_j) \text{ for } j = 0, 1, \dots, 9, 10 \quad (1)$$

where SWB is the level of subject well-being which is the outcome variable with 11 category. The null hypothesis that all coefficients are jointly zero which indicates parallel lines assumption holds. Alternative hypothesis shows that  $\beta$ s differ across binary logit comparisons. According to the test results, it was revealed that some variables violated the parallel lines assumption. Partial proportional odds model was used to solve this problem. This model allows the effect of the predictors that violate the parallel lines assumption to vary across all categories of the ordered outcome variable. The original partial proportional model proposed by Peterson and Harrell (1990) reconstructs the data and determines the interaction between the explanatory variables that violate the parallel lines assumption and the different categories of the ordered dependent variable. The partial proportional odds model proposed by Williams (2006) alleviates the proportional odds assumption by allowing the effect of each explanatory variable to vary across different cut points of the ordered outcome variable without reconstructing the data. The model used can be written as in Eq. (2) which follows methodology of Williams (2006).

$$\Pr(\text{SWB}_i > j) = \frac{\exp(\alpha_j + X M_i \beta M + X K_i \beta K_j)}{1 + \{\exp(\alpha_j + X M_i \beta M + X K_i \beta K_j)\}} \quad (2)$$

where  $i$  shows number of individuals changes between 1 and 20,820;  $\alpha_j$  represents cut points of the model;  $X M$  indicates explanatory variables which do not violate parallel lines assumption in the model,  $M = 1, 2, 3, 4$ ;  $X K$  shows explanatory variables which violate parallel lines assumption in the model,  $K = 5, 6, 7, 8, 9$ ;  $\beta$ s are logit coefficients;  $j$  denotes the category of outcome variable,  $j = 0, 1, 2, \dots, 10$ . This model uses maximum likelihood estimation.

## 4 Data and Results

### 4.1 Data

TURKSTAT 2013 Income and Living Conditions Survey was used in the study. It has information about 12 regions of Turkey according to NUTS-1. Also, it is the latest data set which collects data on individual's subjective well-being and detailed income resources. The data set consists of 33,755 observations. It contains information about the income of 20,820 individuals. Therefore, descriptive analysis covers the whole data set, while the predicted models cover only those who report their income. In the analyzes, in order to investigate the effect of being a beneficiary of

**Table 2** Variables used in the study

Variable name	Definition
Subjective well-being	General life satisfaction that can be valued between 0 (completely dissatisfied) and 10 (completely satisfied)
Health status	1-Very bad; 2-bad; 3-not bad; 4-good; 5-very good
Gender	1-Male; 0-female
Marital status	1-Married; 0-otherwise
Age	15 +
Education level	0-Illiterate; 1-being literate but not graduating from school; 2-primary school; 3-secondary school, vocational secondary school; 4-general high school; 5-vocational or technical high school; 6-college, faculty and above
Household head	1-The reference person is the head of the household; 0-otherwise
Relative income	Absolute income/average income (Rahayu 2016)
Social assistance	1-he/she receives social assistance from the state (child allowance in cash and in-kind, housing allowance, social assistance in-cash and in-kind); 0-otherwise

social assistance on subjective well-being, individuals receiving child benefits in-cash and in-kind, individuals receiving housing allowance and individuals receiving social assistance in cash and in-kind were taken into account. An individual who receives any of these benefits was named as social assistance beneficiary and variable which indicates social assistance beneficiary received a value of 1 if individual receives any of these benefits. It is a binary variable, takes the value 0 for individuals who are not beneficiaries of social assistance. The variables used in the study and their definitions are given in Table 2.

The data set is 44.43% male, 73.80% married, 43.42% household head and 20.93% live in the eastern region. The level of education completed is distributed as follows: 12.60% are illiterate, 8.27% are literate but do not complete a school, 36.64% are primary school graduates, 15.06% are secondary school or equivalent, 8.27% are general high school graduates, 7.43% were technical or vocational high school graduates and 11.73% completed higher education. The age ranges from 15 to 110, and the average age is 43. 5.38% of individuals in the data set are poor. Here, 1.004 TL was used as the equivalent of 1 US dollar in terms of purchasing power parity to determine the poor individuals. Accordingly, poor individuals are determined as those whose income is below 4.3 US dollars according to the daily purchasing power parity. 15.61% of the individuals in the data set are social assistance beneficiaries. 5.70% of the individuals in the data set are not satisfied with their lives at all, however, 7.34% are very satisfied with their lives. To the question about general life satisfaction, 26.35% of the data set answered not bad. This category is the most preferred category among the answers. Average satisfaction value differences were investigated according to poverty status that might be effective in analyzing subjective

**Table 3** Average satisfaction levels by poverty status

Variable name	Poor	Not poor	Not poor/poor
Subjective well-being	5.65	5.22	1.08

**Table 4** Distribution of subjective well-being by income quantiles

Subjective well-being	General	Income quantiles					Share of income not reported
		1	2	3	4	5	
0	5.70	7.52	7.54	6.76	5.00	2.98	5.28
1	1.48	1.87	1.74	1.11	1.45	0.87	1.60
2	3.10	5.28	3.07	2.47	2.36	1.43	3.38
3	6.95	8.93	8.31	7.37	6.60	3.53	6.93
4	7.67	8.81	8.71	8.19	7.13	5.88	7.55
5	26.35	27.67	28.99	29.25	26.93	20.43	25.85
6	12.20	9.92	11.57	12.94	13.54	13.26	12.13
7	12.84	10.33	10.52	11.85	14.20	18.83	12.37
8	12.53	9.75	10.52	10.62	12.94	19.46	12.33
9	3.84	3.36	2.97	1.64	3.69	5.40	4.20
10	7.34	6.56	6.07	6.79	6.15	7.94	8.38
Number of observation	33,755	4164	4202	4126	4196	4132	12,935

well-being in terms of social policy. Average satisfaction levels according to poverty are shown in Table 3.

According to Table 3, individuals who are not poor are generally 8% more satisfied with their lives. Based on this result, quantiles were created to investigate how the subjective well-being of individuals is distributed according to income. For this purpose, first the amount of income is ranked from the lowest to the highest, and then the listed income amount is divided into five equal groups. Table 4 shows the distribution of subjective well-being by income quantiles. According to Table 4, 38.32% of the participants did not report their income. While the group with the lowest income level expressed their subjective well-being level as above not bad (5 points+) at a rate of 39.92%, this rate is 64.89% in individuals with the highest income level. This situation shows the importance of the effect of income status on subjective well-being.

## 4.2 Results

In the analysis, the partial proportional odds model was used to investigate the effect of social assistance on subjective well-being. In the estimated models, observations of 20,820 people who reported their income were used. As the explanatory variables

**Table 5** Results of Brant test

Brant test for model			
Value of the test statistic ( $\chi^2$ )	423.03***		
Brant test for individual variables			
Variable	Value of the test statistic ( $\chi^2$ )	Variable	Value of the test statistic ( $\chi^2$ )
Relative income	51.94***	Social transfer	27.88***
Gender	12.07	Marital status	10.97
Household head	18.68**	Education level	101.22***
Age	6.67	Health status	41.39***
Age square	11.99		

Note \*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001

of the predicted models, relative income, gender, age, marital status, educational level and health status, which have been widely researched in the social assistance and subjective well-being literature, are used. Despite the frequently encountered problem of heteroscedasticity in cross-section data, robust standard errors were calculated in the models. Since the estimation of ordered regression models is based on the assumption of proportional odds ratio, in other words, the assumption of parallel lines, the Brant test was applied in order to decide which ordered regression model should be used (Table 5).

According to the individual Brant test results, the null hypothesis of the parallel lines assumption for the variables of gender, age, age square and marital status could not be rejected. It was concluded that the parameter estimates pass through the same cut-off point. Therefore, it was estimated using a partial proportional odds model, which takes into account that some estimators provide the assumption of parallel lines and some do not. Partial proportional odds model results were given in Tables 6 and 7. Since the log-pseudo likelihood value is statistically significant in the estimated partial proportional odds model, it has been concluded that the model with all independent variables is significant (Wald  $\chi^2_{54} = 2089.43$ ,  $p < 0.001$ ). Since the parallel lines assumption is provided for the variables of gender, age, age square and marital status in the model, the estimators take the same value for each category of the dependent variable. While the head of the household variable was found statistically insignificant in Model 0, Model 1 and Model 2; it was found statistically significant in Model 3 and Model 4 at 1% significance level; similarly significant in Model 5, Model 6, Model 7 and Model 8 at 0.1% significance level and at 5% significance level significant in Model 9. Throughout binary models, the estimated coefficient of the household head variable is in line with expectations and negative. Being the

**Table 6** Results of partial proportional odds model

Variables	Model 0 (Y > 0 versus Y ≤ 0)		Model 1 (Y > 1 versus Y ≤ 1)		Model 2 (Y > 2 versus Y ≤ 2)		Model 3 (Y > 3 versus Y ≤ 3)		Model 4 (Y > 4 versus Y ≤ 4)	
	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio
Relative income	0.2683***	1.3077***	0.2576***	1.2939***	0.2638***	1.3019***	0.2827***	1.3267***	0.2614***	1.2988***
Gender	-0.2112***	0.8096***	-0.2112***	0.8096***	-0.2112***	0.8096***	-0.2112***	0.8096***	-0.2112***	0.8096***
Household head	-0.0359	0.9647	-0.0500	-0.9512	-0.0102	0.9899	-0.1296**	0.8785**	-0.1262**	0.8814**
Age	-0.0592***	0.9425***	-0.0592***	0.9425***	-0.0592***	0.9425***	-0.0592***	0.9425***	-0.0592***	0.9425***
Age square	0.007***	1.0007***	0.007***	1.0007***	0.007***	1.0007***	0.007***	1.0007***	0.007***	1.0007***
Social assistance	-0.2052**	0.8145**	-0.2404***	0.7863***	-0.3641***	0.6948***	-0.4761***	0.6212***	-0.5022***	0.6052***
Marital status	0.2491***	1.2828***	0.2491***	1.2828***	0.2491***	1.2828***	0.2491***	1.2828***	0.2491***	1.2828***
Education level	0.0554*	1.0569*	0.0561**	1.0577*	0.0635***	1.0656***	0.0727***	1.0754***	0.0766***	1.0796***
Health status	0.4832***	1.6213***	0.4697***	1.5996***	0.4909***	1.6337***	0.4716***	1.6025***	0.4228***	1.5262***
Constant	1.7441***	5.7210***	1.5781***	4.8460***	1.1053***	3.0202***	0.6084***	1.8375***	0.2860*	1.3312*

Note \*p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

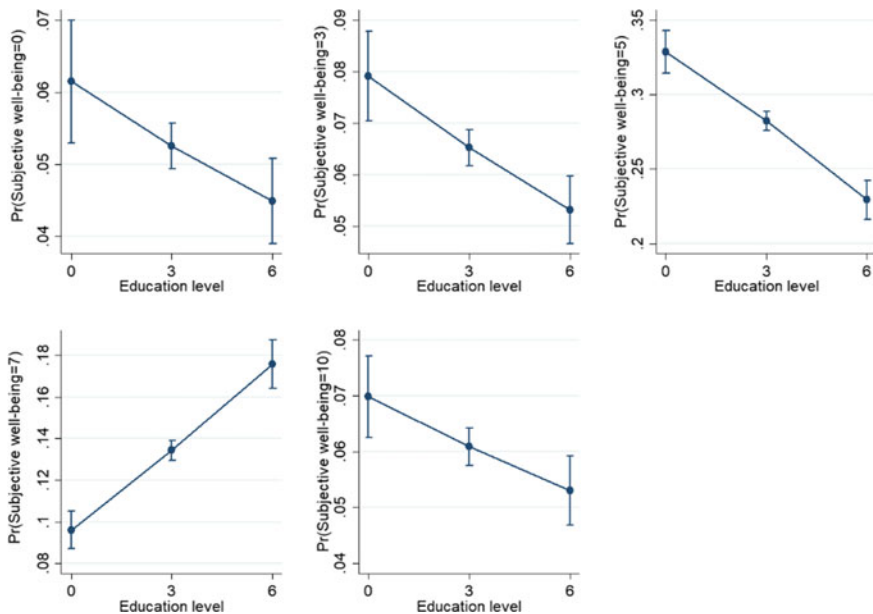
Table 7 Results of partial proportional odds model

Variables	Model 5 (Y > 5 versus Y ≤ 5)		Model 6 (Y > 6 versus Y ≤ 6)		Model 7 (Y > 7 versus Y ≤ 7)		Model 8 (Y > 8 versus Y ≤ 8)		Model 9 (Y > 9 versus Y ≤ 9)	
	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio
Relative income	0.2441***	0.2441***	0.2171***	1.2425***	0.1877***	1.2065***	0.1295***	1.1383***	0.0981***	1.1031***
Gender	-0.2112***	-0.2112***	-0.2112***	0.8096***	-0.2112***	0.8096***	-0.2112***	0.8096***	-0.2112***	0.8096***
Household head	-0.1453***	-0.1453***	-0.1847***	0.8314***	-0.2317***	0.7932***	-0.2519**	0.7773**	-0.1619*	0.8506*
Age	-0.0592***	-0.0592***	-0.0592***	0.9425***	-0.0592***	0.9425***	-0.0592***	0.9425***	-0.0592***	0.9425***
Age square	0.0007***	1.007***	0.0007***	1.007***	0.007***	1.0007***	0.007***	1.0007***	0.007***	1.0007***
Social assistance	-0.4730***	-0.4730***	-0.5125***	0.5990***	-0.5420***	0.5816***	-0.5233***	0.5926***	-0.5759***	0.5622***
Marital status	0.2491***	0.2491***	0.2491***	1.2828***	0.2491***	1.2828***	0.2491***	1.2828***	0.2491***	1.2828***
Education level	0.1229***	0.1229***	0.1132***	1.1199***	0.0728***	1.0755***	0.0023	1.0024	-0.0487**	0.9525**
Health status	0.4048***	0.4048***	0.4166***	1.5168***	0.430***	1.5372***	0.5040***	1.6553***	0.5267***	1.6933***
Constant	-1.0108***	-1.0108***	-1.5108***	0.2207***	-2.0442***	0.1295***	-3.0016***	0.0497***	-3.4225***	0.0326***

Note \*p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

head of the household decreases the odds ratio of being very happy by 14.94% compared to the probability of being in other categories. The variable included in the model in order to investigate the effect of being a social assistance beneficiary on the subjective well-being was found statistically significant at the 0.1% significance level across all binary models. Being a beneficiary of social assistance decreases the odds ratio of having the highest level of subjective well-being by 43.78% compared to the probability of being in other categories. However, a one-unit increase in relative income increases the odds ratio of having better subjective well-being by 10.31%. Education level variable is insignificant in Model 8; significant in Model 0 at 5% significance level; in Model 1 and Model 9 at 1% significance level and in remaining models significant at 0.1% significance level. Education level has a positive effect on subjective well-being. In order to observe the effect of changes in education variable on subjective well-being, the probability of being in categories 0, 3, 5, 7 and 10 of subjective well-being was estimated and shown in Fig. 1.

According to the results, while the probability of subjective well-being in the 0, 3, 5 and 10 categories of those who have completed college, faculty or higher education decreased, the probability of being in the 7th category increased. The increase in the education level of the individual decreases the odds ratio of having the highest level of subjective well-being by 4.75% compared to the probability of being in other categories. However, an increase in an individual’s education level increases the probability of being in the 8th and 9th categories of subjective well-being by



**Fig. 1** The relationship between subjective well-being categories and educational status with adjusted predictions with %95 confidence intervals



0.24% compared to the probability of being in other categories. The health status variable was found significant at the 0.1% significance level in all models, and the sign of the variable was positive in accordance with the expectations. The increase in the health status of the individual increases the probability of being in the highest subjective well-being level by 69.33% compared to the probability of being in other categories. Similarly, the variable of being married was found statistically significant at the 0.1% significance level, and it was found that having the highest subjective well-being compared to the probability of being married in other categories increased odds ratio by 28.28%.

## 5 Conclusions

Since the 1980s, the reduction of the role of the state in the social field has been discussed in developed countries where privatizations, free movement of goods, technology and information, deregulation in competition and neoliberal economic policies are effective in the economic area of the state. This debate stems from the fact that the current social welfare state appears to be an economic burden after the economic crises in the globalizing world under the influence of monetarism. In contrast, social policy applications in Turkey are increasing. Although neoliberal policies are followed in the economy, the interventions of the state to reduce poverty have been increasing in the last decade. These interventions are carried out by the Social Assistance and Solidarity Foundations, which work like a non-governmental organization. However, this situation creates some problems. These problems can be divided into two groups which are individual problems and practical problems. The problems that social assistance can create for the individual are stigma and humiliation, addiction and laziness (Taşçı 2017). On the other hand, the problems that social assistance can create in terms of practice are the recreation of poverty as a result of clientelism, injustice in the form of distribution of social assistance, injustice in the delivery of social assistance, unfairness in the quality of social assistance, unfairness in the continuity of social assistance and increase in demand for social assistance (Kutlu 2015). However, an important point to be noted here is that the problems in terms of practice can also create problems for the individual. In this context, the study investigated how the poverty experienced by the poor in Turkey by using partial proportional odds model. For this purpose, TURKSTAT 2013 income and living conditions which is the latest data set which collects data on individual's subjective well-being and detailed income resources were used. It is also the limitation of the study. Because, the data of 2013 may imperfectly conform with 2021. However, it is a worthy attempt to look at the subject deeply. According to the partial proportional odds model results, the variable included in the model to investigate the effect of being a social assistance beneficiary on subjective well-being was found statistically significant at 0.1% significance level across all binary models. The fact that an individual is a beneficiary of social assistance decreases the odds ratio of having the highest level of subjective well-being by 43.78% compared to the

probability of being in other categories. This situation may be evidence of welfare stigma.

Welfare stigma is an important concept in social management research and is defined as a central problem (Pinker 1971). Stigma marks the person who benefits from prosperity, damages his reputation. In addition, stigma can become a barrier that can prevent the person from accessing social services and an experience that can make them feel degraded. Pinker (1971) used the phrase “it is the most common form of violence in democratic societies” about stigmatization. Stigma is associated with two fundamental problems of social welfare. The first of these is the quality of the services provided. In the quality of services, the attitudes of those who deliver the aid can be humiliating (Spicker 2011). The following participants’ expressions in Kutlu (2015) study show the situations that beneficiaries may encounter in service quality:

Participant 6:

What do you not encounter in the aid distribution. (...)I don’t know, they read names, chaos, crowd, you get it hard, even your food is stolen.(...) It also happens, screaming and calling. We take it under difficult conditions. People are accumulating early in the morning, they are waiting, the truck will come, they will read the list, you are chasing the truck, you are looking for your name. It gives; but it disgraces.

Participant 7:

The people do not stop, they do not wait, the people do not stop, as if there is fear in the people. Since They are worried that their food will go, They will not be able to replace it, it will not come again, They will not be able to buy it, the lame comes there, the blind come there, they come in a misery.

The second problem, closely related to the first problem, is the effect of services on demand. The statements given by the participants of the field study in Kutlu (2018) study show this situation.

Participant Gülseren made the following statements regarding the aids (Kutlu 2018).

It’s not a good thing to get help, you can’t go, when you were ill and go there to get aid, you feel awkward there, you feel like you have lowered yourself a little bit from him. You are human too, I am also human. They offend me, their words are heavy, they either say a word to you, or they say something. You can’t respond as it should.

Participant Muhsin, who did not want to get in the bread queue and buy bread, used the following statements (Kutlu 2018).

No bro. I find two bread wherever they are, don’t get me wrong, Let me tell you as an example, I tell you bro that I’ll buy two bread from here, I take it for the sake of God and go to my house. I can’t get up every day and get in line there. I’m looking at those in the queue, all women, not one man. How can I see myself like that among them? No way, I can’t fit myself. That’s why I never thought of it anyway.

The statements based on field research in qualitative studies confirm that social assistance can have a negative effect on the subjective well-being, as the analysis results show. This is why the study is very important in terms of providing quantitative

evidence how poverty is felt in Turkey. Another important feature of the study is that the study used nationwide representative data set. In this way, we can use the analysis results to make a policy recommendations.

In terms of other variables used in the model, the estimation results are consistent with the subjective well-being literature. The subjective well-being studies in recent years show that relative income is effective on subjective well-being. Here, relativity is that one's income depends on one's expectations, habits, and social comparisons (Diener et al. 1993). According to analysis results, there is a positive relationship between relative income and subjective well-being. Another positive relationship was found between the individual's being married and his subjective well-being. This result confirms the protection/support hypothesis of Coombs (1991), which shows that single individuals who do not have an ongoing relationship with a spouse providing emotional and financial support have more difficulty. Another important result of the study is that the subjective well-being of the person increases as the health condition improves.

The study aims to improve impact of the social policy by providing a basis for further research. It also suggests an integrated approach to social studies which include quantitative and qualitative methods together. Next steps will comprise collect a panel data set to test main hypothesis of the study in order to take into account individual effects and time effects together.

## References

- Akar D (2015) Türkiye'de Primsiz Sosyal Güvenlik Rejimi. *TAAD* 6(21):605–619
- Alper Y (2017) Sosyal Güvenlik. In: Tokol A, Alper Y (eds) *Sosyal Politika*. DORA, Bursa, pp 233–236
- Coombs RH (1991) Marital status and personal well-being: a literature review. *Fam Relat Interdiscip J Appl Fam Stud* 40(1):97–102. <https://doi.org/10.2307/585665>
- Diener E, Sandvik E, Seidlitz L, Diener M (1993) The relationship between income and subjective well-being: relative or absolute? *Soc Indic Res* 28:195–223
- Güneş M (2012) Yoksullukla Mücadelede Sosyal Yardımların Bir Kamu Yönetimi Politikası Olarak Sürdürülebilirliği. *Selçuk Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Sosyal Ve Ekonomik Araştırmalar Dergisi* 24:149–184
- İzgi BB (2008) Türk Sosyal Güvenlik Sisteminde Son Gelişmeler. *Çalışma Ve Toplum* 1:85–106
- Karadoğan E (2018) Sosyal Yardımlar Zararlı (Mı?): Bir Paket Makarna'nın Öz Saygı Düzeyine Etkisi ve Klientalizm Sorgulaması. In: Denizcan Kutlu (ed) *Sosyal Yardım Alanlar Emek, Geçim, Siyaset ve Toplumsal Cinsiyet, İletişim*, İstanbul, pp 207–224
- Killburn K et al (2018) Paying for happiness: experimental results from a large cash transfer program in Malawi. *J Policy Anal Manag* 37(2):331–356
- Kutlu D (2015) Sosyal Yardım Hakkı Tartışması: Türkiye'de Bir Sosyal Haksızlık Olarak Sosyal Yardımlar. In: VII. Sosyal İnsan Hakları Uluslararası Kongresi Bildiri Kitabı. *Sosyal Güvenlik Denetmenleri Derneği*, Denizli, pp 365–382
- Kutlu D (2018) Sosyal Yardımlar Alanlar Konuşuyor. In: Kutlu D (ed) *Sosyal Yardım Alanlar Emek, Geçim, Siyaset ve Toplumsal Cinsiyet, İletişim*, İstanbul, pp 227–400
- Long JS, Freese J (2014) *Regression models for categorical dependent variables using Stata*, 3rd edn. Stata Press Publication, Texas

- Peterson B, Harrell FEJ (1990) Partial proportional odds models for ordinal response variables. *J Roy Stat Soc Ser C (Appl Stat)* 39(2):205–217
- Pinker R (1971) *Social theory and social policy*. Heinemann, London
- Rahayu TP (2016) The determinants of happiness in Indonesia. *Mediterranean J Soc Sci* 7(2):393–404
- Republic of Turkey Ministry of Family Labour and Social Services. <https://ailevecalisma.gov.tr/sygm/programlarimiz/sosyal-yardim-programlarimiz/>. Accessed 14 Jan 2021
- Republic of Turkey Ministry of Family Labour and Social Services. <https://ailevecalisma.gov.tr/raporlar/yillik-faaliyet-raporlari/>. Accessed 14 Jan 2021
- Spicker P (2011) Stigma and social welfare. Creative Commons, USA. <https://www.spicker.uk/books/Paul%20Spicker%20%20Stigma%20and%20Social%20Welfare.pdf>
- Taşçı F (2017) *Sosyal Politika Ahlakı*. Kaknüs Yayınları, İstanbul
- Williams R (2006) Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata J* 6(1):58–82

# Formation of a Fishing and Aquaculture Cluster as a Tool for Regional Competitiveness



María Francisca Peñaloza-Talavera, Jaime Apolinar Martínez-Arroyo, and Marco Alberto Valenzo-Jiménez

**Abstract** The purpose of this research is to analyze the viability of the emergence of a fishing and aquaculture cluster in the state of Michoacán, with the aim of becoming a tool that promotes the regional Competitiveness of the territory. To determine the feasibility of the formation of an agglomeration of companies, in this work the methodology proposed by (Fregoso in Factores determinantes en las asociaciones para formar clústers industriales como estrategia de desarrollo regional. Tesis Doctoral, Instituto Politécnico Nacional, México, 2012) is used, where the use of coefficients is considered to determine mathematically if the emergence of a cluster in the region. To apply the coefficients, we start from the proximity theory and the number of companies in the sector and in the industry, the number of workers in the sector and in the industry and the employed population in the industry are used as essential factors for the calculation. The information is collected from the INEGI Economic Census INEGI (Recuperado el 10 de 09 de 2020, de, 2019) and, by substituting the data in the coefficient formulas, it is concluded that, given the state's agricultural and specifically fishing vocation, the emergence of a fishing cluster that promotes the Regional Competitiveness is a feasible possibility in the Infiernillo, Costa, Tierra Caliente, Pátzcuaro, Cuitzeo and Lerma—Chapala regions of Michoacán.

**Keywords** Cluster · Regional competitiveness · Company · Fishing · Coefficient

---

M. F. Peñaloza-Talavera · J. A. Martínez-Arroyo (✉) · M. A. Valenzo-Jiménez  
Faculty of Accounting and Administrative Sciences, The Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico  
e-mail: [jmartinez@umich.mx](mailto:jmartinez@umich.mx)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_16](https://doi.org/10.1007/978-3-030-85254-2_16)

267

## 1 Introduction

Nowadays, Small and Medium-sized Enterprises (SMEs) play vital roles in most countries involving various aspects of the economy, including manufacturing and services. Indeed, these enterprises are major providers of employment, evolution, and innovation, as well as the pioneers in novel technology inventions (Babkin et al. 2013). Accordingly, the development of SMEs facilitates domestic development of the country and accelerates industrial growth. Today, planning for the development of SMEs, based on a clustering approach is considered as a method to achieve developmental goals in many countries (Karaev et al. 2007). In this way, clusters are a form of integration of economic entities, sectorial institutional structures, regional institutional organizations, based on mutually beneficial cooperation, technological exchange, qualification, which creates competitive advantages for the whole structure and its parts, contributes to the development of production, competitiveness, transaction costs reduction and effective access to foreign markets (Romanova et al. 2018).

This means that clusters are a suitable method for creating competitive advantage, not only for enterprises of the same cluster but also for the country on which the clusters are based (Shakib 2020). In such a way, the enterprises are influenced by the evolution occurring inside a cluster. However, the effects of many factors should be considered in a clusters' developmental plan and it is important to determine such factors, or variables (Danesh et al. 2017).

The development of cluster structures boosts regional and National economic processes, which has a positive effect on the investment attractiveness and socio-economic potential of the region and leads to the creation of Regional Competitiveness with new enterprises and jobs (Isaksen 2018).

It should be noted, In the twentieth century, clusters began to be considered the most important factor in regional development (Kozonogova, Elokhova, Dubrovskaya, & Goncharova). Regions with developed cluster structures are more competitive; clusters are a foothold for successful regional economies. The aggregation of enterprises and organizations into cluster makes it possible to increase their effectiveness (Kudryavtseva et al. 2020).

In this way, micro, small and medium-sized enterprises (MSMEs) benefit from these type of strategies since they are the ones that face the greatest difficulties, due to their size, difficulties ranging from lack of access to external sources of financing, unprofessionalized accounting, low level of investment in innovation, up to the lack of access to adequate sources of information for making rational decisions.

For these types of companies, the cooperation and networks they establish with others of the same size or with large firms is a strategy that allows them to take advantage of the competitive advantages of the companies with which they are related. When the established agreements include a large number of companies established in a common geographic site, a business fabric called a cluster is generated (Saavedra 2012). Briefly, the present study was carried out to identify the viability for the formation of a fishing cluster in the State of Michoacán, Mexico, to solve the problem

associated with the lack of development in MSMEs, therefore, the purpose of this research is to analyze the viability of the emergence of a fishing and aquaculture cluster in the state of Michoacán, with the aim of becoming a tool that promotes the regional Competitiveness of the territory. Finally, the proposed model is dynamic and comprehensive, since it is capable of evaluating different economic scenarios and activities to evaluate the potential possibility of forming a cluster.

The term “cluster” became popular in the late 1980s and is currently used to refer to one of the tools in effective regional development. Today, clusters are studied by scientists around the world. During the twentieth century, a lot of research, within the theoretical frameworks for economic growth and development, focused on ways of optimizing the locations of enterprises and industries in terms of transport and resource constraints (Kudryavtseva et al. 2020). Sustainable cluster development is widely discussed in the scientific literature; it focuses on the idea that clusters can be central to not only a region’s economy, but also its sociology, ecology, and innovation spheres (Chen et al. 2020). For example, some studies only estimate the cluster’s economic performance and make conclusions that focus on its efficiency (Putri et al. 2016). Alike, many researchers have paid contributed to the study of innovative clusters, defining the research sector’s large role in cluster development (Wiratmadja et al. 2016).

The clusters are productive articulations of companies that arise from the identification of the productive vocation of a territory. One of the advantages offered by this type of network is that it gives companies the ability to generate economies of agglomeration that lead them to increase the competitive capacities of the region, making efficient use of resources and thus achieving high levels of productivity and competitiveness (Bonales et al. 2016). In fact, according to the Global Cluster Initiative Survey conducted on 500 clusters around the world, 85% of cluster initiatives have improved their competitiveness, 89% have helped the cluster to grow, and 81% have met their goals. In contrast, only 4% have been disappointing and did not lead to much change (Otsuka and Ali 2020).

So that, three main reasons can be noted for the existence of these agglomeration forces. The first is access to efficient labor markets with a specialized labor supply and easier matching processes between workers and firms. The second is the reduction of transaction costs due to the spatial integration of different steps of the production process. The third and main argument for the analysis is the existence of knowledge spillovers via proximity interactions (Hassine and Mathieu 2020).

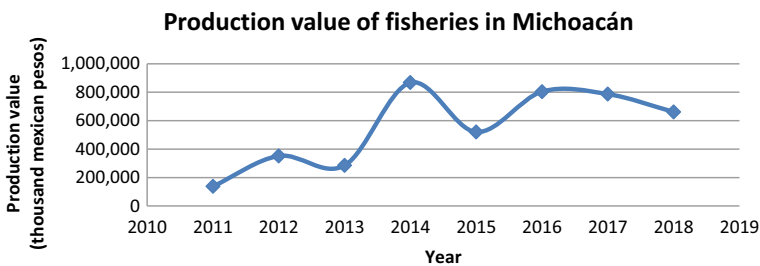
The purpose of this study is to analyze the feasibility of a cluster emerging from the Michoacán fishing sector, in order for the companies engaged in this activity to find an opportunity to improve their possibilities and their ability to compete in a global market, using a tool that encourages regional competitiveness, thereby seeking not only to improve productivity but also the quality of life of the population of the territory. The cluster represents the community of industry and related companies on the basis of cooperation and competitive links. This community is capable of mutually reinforcing competitive advantages based on synergy. Sectorial clusters are inseparably linked. Cluster, as a stable network of interrelated economic entities, has a production potential higher than the sum of the potentials of its participants. This

is achieved as a result of network cooperation and within the network competition. The cluster approach allows to effectively solve the problem of the competitiveness of various levels of economic entities in the National Economy, to create an effective basis for implementing the state policy of economic growth, to increase the effectiveness. To strengthen cooperation and integration of large and small businesses, to accelerate solution of social problems of regions and the National Economy as a whole (Romanova et al. 2018).

It is necessary to look for alternatives that promote fishing activity in Michoacán since during the last two years, the lack of strategies to promote the sector has led to a considerable decrease in activity as shown in Fig. 1. Based on the above. This research considers it pertinent to analyze the situation of the fishing activity in Michoacán in order to expose the possibility of creating a fishing cluster in the state and thereby show the opportunity to improve the competitiveness of the regions studied. In this sense, the question that guides this work is: What regions of the state of Michoacán have the capacity to form a fishing cluster? Based on the question, the main objective of the work is to: Identify the regions of the state of Michoacán that have the capacity to develop as a fishing cluster. In order to answer the research question and fulfill the objective of the work, in this work the theoretical foundation of the study is first presented. For this, concepts such as competitiveness, regional competitiveness and the cluster are addressed.

Competitiveness is a variable widely analyzed today because its promotion not only has positive effects on macroeconomic indicators, but also on the quality and living conditions of citizens (Sarmiento 2019; Domínguez y Gutiérrez 2017; Ordóñez Tovar 2011; Vázquez y Reyes, 2013). According to recent studies, competitiveness at the nation and region level depends largely on the way in which companies take advantage of the natural and created competitive advantages that the territory offers, such as the availability of natural resources, the climate, the culture, institutions, policies, infrastructure, etc. (Granados et al. 2016; Sarmiento et al. 2015; Ruiz-Velazco 2015; Sarmiento 2019; Morales de Llano 2014; Burbano et al. 2011; Quero 2008).

Despite the large number of studies carried out on the subject of competitiveness, there is no single and generalized definition of this term, because, as pointed out Saavedra y Milla (2012) competitiveness is an extremely complex concept, which



**Fig. 1** Source: Own elaboration using data from SIAP



can be studied from different approaches and perspectives. Given the breadth of the concept of competitiveness, some of the main definitions of the term that are useful to understand the concept under the context of analysis of this research are presented below.

One of the most cited definitions of competitiveness is the one proposed by the World Economic Forum, an institution that defines it as “the set of institutions, policies and factors that determine the level of productivity of a country” (World Economic Forum 2010, p. 4). Based on this definition, productivity is a complementary variable to competitiveness and it is necessary to clarify that they do not have the same meaning. In that sense, for there to be competitiveness in a nation, it is required that it have solid bases of productivity. In addition to generating economic benefits, competitiveness based on productivity produces significant changes in the level of prosperity and well-being of citizens, which is why it is considered a necessary element for the development of a country (Suárez 2005).

For Porter (1991), the competitiveness of a nation is associated with the productivity of its companies, which depends on the efficient use of natural, human and capital resources. Thus, competitiveness is associated with the ability of swarms or agglomerations of companies to use their human, natural and capital resources productively. (Ibarra et al. 2017; Arana y Ballesteros 2016; Valencia y Zetina 2017). Under the same approach, La barca (2007, p. 161) points out that competitiveness refers to the “possibility that its citizens have to achieve a high and growing standard of living” and this possibility depends on the “productivity with which national resources are used, the product per unit of work or capital used”. In that sense, competitiveness can be achieved when existing economic units achieve higher productivity. The perspective of Labarca (2007) is very consistent with the definition proposed by the World Economic Forum (2010) and with the ideas of Porter (1991); All three approaches consider that the variables productivity and quality of life are fundamental for competitiveness.

According to Porter (1999) competitive advantage is created and sustained in localized processes, that is, at the territorial level or what is the same at the regional level, and the main objective of competitiveness is to increase productivity. And thus lead to improving the quality of life of the population residing in a specific region (Krugman 1994). At the territorial level, regional competitiveness helps to subsidize competitiveness at the company level. And it is that, finally, who competes and acts in the market is the company as such. Upon emerging from the local context, the strategies and processes developed by the companies are clearly influenced by the culture of the territory, in that sense, their actions are largely supported by socialized processes or explicit collective action of a territorial nature. For this reason, it is common for territories to generate networks and business agglomerations (Camagni 2005).

In sum, the concept of competitiveness is closely related to the productivity and well-being of citizens in a given territory. Productivity is associated with the performance of natural, human and capital resources possessed by the economic units of a country, a region or a productive system. The way in which companies located in a specific territory take advantage of their resources has a direct impact on the quality

of life of the population that belongs to a region. Under this logic, it is possible to affirm that competitiveness is widely related to the productivity of production systems, which are generally located in geographic regions with a specific productive vocation and that give rise to the so-called agglomerations of companies or clusters.

Clusters are geographic concentrations of companies and institutions that share a common interest and are considered a tool to boost the regional competitiveness of strategic sectors; thanks to the proximity, the process of exchange of information, knowledge and technology between economic units is facilitated (Arana y Ballesteros 2016). In the last decade, the success of industrial districts in developed countries has stimulated the creation of the cluster approach to explain the agglomeration of small companies in developing countries. (Bada y Rivas, 2010; Rabellotti y Pietrobelli, 2005). Therefore, when forming a cluster, one of the most important criteria is the maximization of the potential of the cluster structure. However, the objective function may be not necessarily the criterion for the formation of the system. Corporate processes and self-organization in open and non-equilibrium systems lead to their dynamic stability (Romanova et al. 2018).

Cluster is defined as a geographically approximate concentration or group of interconnected companies and associated institutions in a common interest in a specific economic and strategic sector and linked to complement each other (Porter 1999). The analysis of clusters and their importance as a factor of competitiveness can be understood from different theories that explain their emergence. Among the theories that try to explain the emergence of clusters, five main ones can be pointed out: the theory of location and economic geographic, the theory of backward and forward linkages, the theory of the interaction of industrial districts, the theory of Michael Porter's competitive advantage and the theory of economic growth based on consumer goods. Each of the aforementioned theories explains the causes that give rise to the emergence of clusters in the territories and, all of them "share the notion that the competitiveness of each company is strengthened by the competitiveness of the group of companies that make up the group" (Bonales et al. 2016).

The theory of location and geographic economic to evaluate the feasibility of the formation of a cluster, explains the reasons why productive activities are located in specific geographic areas and are not randomly distributed in the territory (North 1995; Krugman, 1995). This indicates that the costs of distance and transport affect the location of the companies; Under this logic, some of the activities are located especially in territories close to natural resources and other activities are located in areas close to the markets that will supply. In this sense, the variable that explains the origin of a cluster under this approach is proximity to resources and markets (Bada y Rivas, 2010).

## 2 Methodology

This research uses the theory of location and geographic economic to identify a potential fishing cluster in Michoacán. Although there are a large number of studies and methods for analysis, most of them focus on the qualitative characteristics of the territories and, those few studies that use quantitative data generally do not include in their formulas the relationship between work and the number of companies in the territory as key indicators for the identification of a true conglomerate of productive units (Fregoso 2012).

This is extremely important, because a large part of the economic activities that are practiced in Mexico are carried out by contract, in that sense there are sectors where one or two companies control the activity and employ almost all workers in the sector. When this situation exists and it is sought to identify a cluster with the traditional formulas that do not use the number of companies within the analysis, the result may be wrong by showing the existence of a potential cluster by concentrating a high number of workers in the sector, forgetting that few companies employ them; in those cases, when those signatures disappear, the potential cluster also disappears. Sforzi (1987), pointed out that a necessary condition for the correct identification of an industrial district is that the employed personnel are employed in a greater number of firms and not grouped in an oligopoly (Sforzi 1987).

Therefore, it is important to use a coefficient that analyzes the concentration of employment in an activity and that, in turn, is capable of relating the number of workers with the number of potential companies in the cluster. One way to achieve the above and geographically delimit a territory is through the mathematical model exposed by Fregoso (2012), which consists of geographically identifying a region that is a cluster by relating the number of workers with the number of companies located in the territory using ratios and quotients. The indicated method is based on the geographic delimitation variable, understood as the concentration and proximity between the companies located in a geographically delimited region. For this type of study, it is necessary that concentration and proximity be defined in quantitative terms. For this reason, the size of the companies in terms of workers per company, personnel employed in the sector and in the industry and, the number of companies or economic units in the sector and industry.

The model exposed by Fregoso (2012), It is based on the theory of location and economic geography and is mainly based on the location coefficient, also called the specialization ratio, from which two coefficients are derived that measure the relationship between employment and the number of companies. The location coefficient represents the relationship between the share of sector  $i$  in region  $j$  and the share of the same sector in the national/regional total. This coefficient is then a way to measure the relative specialization in a territory, in that sense it works as a measure of regional concentration of a certain activity or economic sector. To evaluate this indicator, various variables such as Gross Domestic Product, Value Added and Employment can be used. In this case, it is considered appropriate to use this last variable, employment.

The use of the employment variable to calculate the location index has its theoretical basis in the center-periphery model exposed by Krugman (1991). According to this author, companies and workers tend to concentrate in delimited geographical areas because there is a strong relationship between labor mobility, growing company returns, and trade costs. In this sense, with the employment values, the location coefficient can be calculated, weighing the weight of the number of workers in the activity studied with respect to the total in the region and comparing with the proportion that the same region has at the national level. The values generated by the coefficient define the viability of the emergence of a cluster specialized in the activity studied within a geographically delimited territory (Lira y Quiroga, 2009).

According to Soler (2000), the original mathematical expression of the localization or specialization coefficient is as follows:

$$CE_{ij} = \frac{X_{ij}/X_j}{X_i/X_T}$$

where:

$X_{ij}$  = value of employment in sector i in territory j.

$X_j$  = total value of employment in territory j.

$X_i$  = value of the employment variable for sector i.

$X_T$  = total value of employment in the entire reference territory.

Based on the above proposal formula, Fregoso's (2012), To geographically delimit a territory, the evaluation includes not only the employment factor, but also the number of companies in the territory, since only in this way is it possible to verify that there is an agglomeration of companies in the area and not only a concentration of employees in some business accounts. Based on these adjustments, Fregoso's model (2012) is expressed in three coefficients: cluster coefficient (CC), coefficient of economic unit per work in the sector (CULS) and coefficient of economic unit per labor in the industry (CULI).

The cluster coefficient (CC) is based on the specialization coefficient and its purpose is to determine whether in a territory there is specialization in any economic sector based on the employment variable. In this sense, this coefficient is calculated by weighting the weight of the number of workers in the sector with respect to the total number of employees in the industry and comparing said weighting with the proportion that the same industry maintains at the regional level. The formula for the cluster coefficient (CC) is as follows:

$$CC = \frac{TTS/TTI}{TTI/PTE}$$

where:

TTS = Total labor value in the sector. Expressed in number of people.

TTI = Total labor value in the industry. Expressed in number of people.

PTE = value of the working-age population in the region. Expressed in number of people.

The interpretation of the CC is the same as that performed for the specialization coefficient. The CC formula can yield three different results: less than, equal to or higher than one. When the value of the coefficient is equal to or greater than unity, it indicates that there is specialization in the economic activity analyzed in the region. In this sense, a value greater than one will indicate that there is a potential cluster in the studied region since most of its population of working age works in industry. In addition to the above, it should be added that the higher the CC value, the greater degree of specialization there is in the region. When the result of the cluster coefficient is less than one, it is understood that there is no specialization in the economic activity studied and, therefore, the cluster is not a relevant possibility in the region since most of the working-age population is working in another industry (Lira y Quiroga 2009; Fregoso 2012; Soler 2000).

With the results of the CC value, it is possible to affirm or reject the possibility that a potential cluster exists in a region. However, Fregoso's model (2012), differs from the rest of traditional models because in addition to seeking to establish the viability of a cluster with the CC, it includes the mathematical analysis of two reasons that indicate whether the condition that Sforzi (1987) had already indicated as necessary for an industrial district to exist is fulfilled, this is that the employed personnel are employed in a large number of firms and are not agglutinated in a few companies. For this analysis, the economic unit coefficient per labor in the sector (CULS) and the economic unit coefficient per labor in the industry (CULI) are used, whose formula is the following:

$$CULS = \frac{TTS}{UES} > CULI = \frac{TTI}{UEI}$$

where:

TTS: Total labor value in the sector. Expressed in number of people.

TTI: Total labor value in the industry. Expressed in number of people.

UES: Value of economic units in the sector. Expressed in number of companies.

UEI: Value of economic units in the industry. Expressed in number of companies.

The economic unit coefficient per work in the sector calculates the number of employees per company, in that sense it is a comparison of the number of employees in the economic activity with cluster potential with the number of companies in the sector (CULS). The calculation of the CULS is necessary because it is finally a measure that helps to establish whether the sector has an agglomeration of companies

and employees in the region or if this agglomeration is only of employees in a few companies. The same expresses the coefficient of economic unit for labor in the industry (CULI), it indicates if the employees are agglutinated in a few companies within the industry or if they are distributed in a greater number of companies.

The numerical analysis of these coefficients must meet the condition expressed in the formula:  $CULS > CULI$ . When the condition is met, it is confirmed that there are enough companies in the sector to absorb employment in the place where the economic activity under analysis is concentrated, this also shows the importance of the number of companies within the sector and corroborates that the agglomeration in the region does not it is only of workers but also of economic units. When the condition of the formula is not fulfilled, that is, when  $CULS < CULI$ , it is verified that mathematically it is not possible to absorb the employment of the place with existing companies in the sector, so that the cluster is not a viable option for the region because there are not enough economic units that truly form an agglomeration (Fregoso 2012).

The proper functioning of the model described here is subject to the fulfillment of some technical conditions. The first condition that must be met before starting the application of the model has to do with a previous analysis of the activity, it is essential to identify if there is an agglomeration of companies in the study region, since if there is no indication of agglomeration no matter how small, the evaluation of the CC would not make sense. A second condition that must be met is to identify, prior to the application of the model, the territory where the cluster will be geographically delimited, since if the area is not delimited, it is not possible to correctly define the values of the variables to be used.

The third condition that must be met is to agree with the definition of the cluster concept since there are many conceptualizations and each one of them highlights different important elements for the analysis. And, the fourth condition for the correct operation of the model has to do with the information necessary for the calculation of the coefficients, it is required that there are updated databases that contain the values of the variables used in the model; Specifically, it is necessary to have current information on the number of workers in the sector and in the industry, the number of companies in the sector and in the industry, and the number of people of working age in the region (Lira y Quiroga 2009; Fregoso 2012; Soler 2000).

The model described can be applied and correctly evaluate the mathematical viability of the emergence of a cluster in practically all the productive activities belonging to the three economic sectors (primary or agricultural, secondary or industrial and, tertiary or services) as long as the four conditions indicated above are met. The fourth condition is especially important since if the information is not current, the result obtained may reflect a situation that no longer represents the real conditions of the region. For this reason, in the case of Mexico, the lack of updated databases makes it difficult to apply the model in key activities for the country such as agriculture, livestock and mining; For these activities, the latest available database presents information for the year 2007 (Censo Agrícola et al. 2007, INEGI). The activities belonging to the industrial sector and the services sector in Mexico show a different panorama for the model, since the databases that contain information regarding these activities are updated every 5 years (Economic Censuses).

In the case of this study, the model is applied to the Michoacán fishing sector and two economic activities of the same are analyzed and evaluated: Fishing and Aquaculture, activities 11,411 and 11,251, respectively, in the North American Industry Classification System 2018 (SCIAN). The data used to apply the method is updated and has been obtained from the National Institute of Statistics and Geography (INEGI), specifically from the information available in the 2019 Economic Censuses. The analysis is carried out at the national level, at the state level and at the regional level within the state of Michoacán. To determine the value of the factors in each region, a summation of the individual value of each municipality is made within the region to which it belongs. In total, the ten administrative regions of the state are studied, in which the 113 municipalities that compose it are contemplated.

### 3 Results

This section presents the results of the calculation of the coefficients of clusters CC, CULS and CULI in order to verify the viability of the emergence of a fishing cluster in Michoacán. Firstly, Table 1 presents the results obtained from the analysis of the activity 11,411 corresponding to Fishing; the study is applied at the national, state and regional levels.

Based on the statistics obtained in the Economic Census (INEGI 2019) and the application of the equations to quantitative data determines that the emergence of a fishing cluster at the national level is not viable, as well as in the Meseta, Tepalcatepec, Oriente and Bajío regions. The analysis of the quotients using data from fishing in Michoacán shows that Michoacán is a territory with a productive fishing vocation with cluster potential, its CC is greater than 1 and the comparison of CULS and CULI confirms the existence of companies close to the natural resource studied here and it has been confirmed that the number of existing companies is sufficient to absorb the labor available for the sector at the state level. For this reason, it is accepted that in Michoacán there is the possibility of forming a fishing cluster that contributes to improving the productivity of the companies involved in terms of the use they make of available resources.

Within the Michoacan territory, the regions identified with the possibility of forming a fishing cluster are the following six: Infiernillo, Costa, Tierra Caliente, Pátzcuaro, Cuitzeo and Lerma Chapala. These six regions present a CC greater than 1 and the comparison of CULS and CULI verifies that the conglomerate of companies related to the activity in each region are sufficient to occupy the available labor in the territory and are close to the source of the resource from where they obtain the product and to the market where they commercialize it (see Table 1). Based on the result, it is stated that the fishing cluster can be an adequate tool to promote the competitiveness of the indicated regions, since through the linking of companies it is more feasible to improve the productivity of organizations, with which it can improve the well-being of citizens, well-being that can be promoted from the same

**Table 1** Fishing sector

Region	TTI	TTS	UEI	UES	PTE	CC	CEULS	CEULI	Reasons
National	233,554	179,478	24,372	19,627	27,132,927	89.27	9.14	9.58	False
Michoacán	8632	5861	842	420	779,733	61.33	13.95	10.25	True
Infiernillo	1454	1282	181	157	18,449	11.18	8.16	8.03	True
Costa	1664	1549	114	95	57,136	31.96	16.30	14.59	True
Tierra Caliente	217	131	41	13	20,573	57.23	10.07	5.29	True
Pátzcuaro	587	449	122	67	33,816	44.06	6.70	4.81	True
Meseta	901	1	49	1	119,129	0.14	1	18.38	False
Tepalcatepec	67	12	25	7	53,549	143.14	1.71	2.68	False
Oriente	522	12	145	12	62,407	2.74	1	3.6	False
Cuitzeo	1424	1063	83	36	229,941	120.53	29.52	17.15	True
Bajío	251	9	25	9	70,014	10.0	1	10.04	False
Lerma—Chapala	842	470	57	30	115,576	76.61	15.66	14.77	True

Source Own elaboration using data from Economic Census, 2019 (INEGI 2019)



territory with the grouping of companies around a fishing cluster that contributes to using the available resources more efficiently.

When performing the mathematical analysis in the aquaculture sector at the regional, state and national levels, it is concluded that this activity does not meet the values required by the coefficients to consider the possibility of the emergence of a cluster, as can be seen in Table 2. In Michoacán, the aquaculture activity is in full growth and the potential for its development is wide given the conditions of the territory. Therefore, when applying the CC, the value turned out to be greater than 1, however the possibility of the emergence of a cluster could not be confirmed with the values of CULS and CULI. In this sense, at this time, the formation of a cluster at the Michoacán level is not considered mathematically feasible. In each of the ten regions that make up the state, the value of the CC and/or the comparison of the result of the CULS and CULI, show that there is no potential for an aquaculture cluster in the territory.

## 4 Discussion

Beneficial effects of clusters, manifest themselves through improvements in the quality and efficiency of labor. By improving competitiveness, the economy and the transformation of the industry toward a more productive growth pattern, the main strategy would be the clusters as they play an important role. Although many studies confirm that industrial conglomerates are beneficial for the performance of companies, some issues remain controversial, such as the economic impact of conglomerates in traditional industries and developing countries. The research results confirms the existing theory and complements with new contributions in show the viability of creating a fishing cluster especially now that would be after the crisis experienced.

Alike, higher levels of interactions increase the rate of learning by doing thus increasing labor productivity. In a parallel development, cluster effects enable the release of resources that can be used to attract new, and upgrade the quality of existing, human capital.

The findings of this study show that it is feasible to form a fishing cluster that contributes to improving the productivity of the companies involved in addition to creating a positive impact in the region.

The analysis of quantitative data related to fishing and aquaculture in Michoacán, reveals that mathematically it is possible the emergence of clusters related to the activity. Specifically in the fishing activity, it was found that six Michoacan regions are susceptible to the formation of a cluster that promotes regional competitiveness through the linkage of fishing companies. By studying the theoretical information and comparing it with the empirical data, it is possible to affirm that business clusters are a viable option to facilitate the productivity of fishing units. With strategies such as the cluster, the decrease in fishing production can be stopped and, in this way, reverse the negative effect that the situation is generating on the levels and quality

**Table 2** Aquaculture sector

Region	TTI	TTS	UEI	UES	PTE	CC	CEULS	CEULI	Reasons
National	233,554	33,768	24,372	3666	27,132,927	16.79	9.21	9.58	False
Michoacán	8632	1179	842	342	779,733	12.33	3.44	10.25	False
Infiernillo	1454	63	181	46	18,449	0.54	1.36	8.03	False
Costa	1664	115	114	16	57,136	20.37	7.18	14.59	False
Tierra Caliente	217	86	41	26	20,573	37.57	3.30	5.29	False
Pátzcuaro	587	138	122	55	33,816	13.54	2.50	4.81	False
Meseta	901	31	49	25	119,129	4.54	1.24	18.38	False
Tepalcatepec	67	12	25	6	53,549	143.14	2	2.68	False
Oriente	522	344	145	131	62,407	78.78	2.62	3.6	False
Cuitzeo	1424	165	83	42	229,941	18.71	3.92	17.15	False
Bajío	251	10	25	10	70,014	11.11	1	10.04	False
Lerma—Chapala	842	12	57	8	115,576	1.95	1.5	14.77	False

Source Own elaboration using data from Economic Census, 2019 (INEGI 2019)

of life of the population of the territory. Finally, competitiveness not only seeks to generate productivity but also to improve people's quality of life.

In the case of aquaculture, it was found that the creation of a cluster over it is not feasible because the existing companies are not enough to monopolize the available labor. This situation found is not conclusive due to the limited statistical information found on said activity. As the data does not reflect the total number of aquaculture units that exist in the municipalities the exact amount of labor in the sector, it is not possible to affirm that aquaculture is irrelevant in the territory, this mainly because the data presented in COMPESCA reports (Comisión de Pesca del Estado de Michoacán, México) reveal that it is an activity with great development potential and that every day it involves more productive units, in fact, in the latest report of said institution Michoacán is identified as the second largest aquaculture producing state at the national level. Given this situation, it is considered appropriate to point out that the results obtained with the application of the coefficients on aquaculture activity are not conclusive; in future research the activity can be analyzed in greater detail and define whether or not it is possible to develop a cluster.

## References

- Arana O, Ballesteros A (2016) Los clúster como herramienta para dinamizar la competitividad. *Dictamen Libre* 18:83–93
- Babkin A, Kudryavtseva T, Utkina S (2013) Formation of industrial clusters using method of virtual enterprises. *Proc Econ Finance* 5:68–72
- Bada L, Rivas L (2010) Los clústers agroindustriales en el estado de Veracruz. *Investig Adm* 105:73–100
- Bonales J, Martínez J, Valenzo M (2016) Modelo competitivo de clusters de empresas exportadoras del estado de Michoacán. *Memoria del IX Congreso de la Red Internacional de Investigadores en Competitividad*
- Burbano E, González V, Moreno E (2011) La competitividad como elemento esencial para el desarrollo de las regiones. *Una mirada al Valle del Cauca. Revista Gestión y Desarrollo* 8(1):51–78
- Camagni R (2005) *Economía urbana*. Barcelona
- Chen X, Wang E, Miao C, Ji L, Pan S (2020) Industrial clusters as drivers of sustainable regional economic development? An analysis of an automotive cluster from the perspective of firms' role. *Sustainability* 12(7):1–22
- COMPESCA. (s.f.). Recuperado el 05 de 10 de 2020, de <http://compesca.michoacan.gob.mx/michoacan-con-gran-potencial-pesquero-compesca/>
- CONACYT. (s.f.). Sistema Nacional de Investigadores. Recuperado el 08 de 07 de 2019, de <https://datos.gob.mx/busca/dataset/sistema-nacional-de-investigadores/resource/3ecd4bae-63a2-43c5-82e0-8619175e29a2>
- Danesh SM, Toloie EA, Alborzi A (2017) Identification and evaluations of factors involved in industrial clusters development, applying fuzzy DEMATEL. *Int J Appl Manag Sci* 9(2):135–152
- Domínguez J, Gutiérrez A (2017) La competitividad y el desarrollo económico de las empresas exportadoras de orégano seco en la región Tacna. *Universidad San Ignacio de Loyola, Facultad de Ciencias Empresariales, Lima*
- Fregoso G (2012) Factores determinantes en las asociaciones para formar clústers industriales como estrategia de desarrollo regional. *Tesis Doctoral, Instituto Politécnico Nacional, México*

- Granados H, Giraldo Ó, Acevedo N (2016) Promoción de la competitividad y el desarrollo territorial en los municipios del Valle de Aburrá. *Semestre Económico* 19(40):93–116
- Hassine HB, Mathieu C (2020) R&D crowding out or R&D leverage effects: an evaluation of the french cluster-oriented technology policy. *Technol Forecast Soc Chang*. <https://doi.org/10.1016/j.techfore.2020.120025>
- INEGI (2007) Censo Agrícola, Ganadero y Forestal. Recuperado el 15 de 10 de 2020, de <https://www.inegi.org.mx/programas/cagf/2007/>
- INEGI (2019) Recuperado el 10 de 09 de 2020, de <https://www.inegi.org.mx/programas/ce/2019/>
- INEGI CE (2019) Censos Económicos. Recuperado el 10 de 12 de 2020, de <https://www.inegi.org.mx/programas/ce/2019/#Tabulados>
- INEGI (s.f.) PIB y cuentas nacionales. Recuperado el 05 de 08 de 2019, de <https://www.inegi.org.mx/temas/pib/>
- Isaksen A (2018) From success to failure, the disappearance of clusters: a study of a Norwegian boat-building cluster. *Camb J Reg Econ Soc* 11(2):241–255
- Karaev A, Koh SC, Szamosi LT (2007) The cluster approach and SME competitiveness: a review. *J Manuf Technol Manag* 18(7):818–835
- Kozonogova E, Elohova I, Dubrovskaya J, Goncharova, N (s.f.) Does state cluster policy really promote regional development? The case of Russia. In: IOP conference series. Materials science and engineering, vol 497, issue 1, p 120
- Krugman, P. (1994). Competitiveness. A dangerous obsession. *Foreign Affairs* 73(2):28–44
- Krugman P (1995) Development, geography and economic theory. MIT Press, Cambridge
- Kudryavtseva T, Skhvediani A, Ali M (2020) Modeling cluster development using programming methods: case of russian arctic regions. *Entrepreneurship and Sustainability Issues* 8(1):150–176
- Labarca N (2007) Consideraciones teóricas de la competitividad empresarial. *Omnia* 13(2):158–184
- Morales de Llano E (2014) La dimensión territorial de la competitividad. *Economía y Desarrollo* 151(1):71–84
- North D (1995) Location theory and regional economic growth. *J Polit Econ* 6:56–67
- Ordóñez Tovar J (2011) ¿Competitividad para qué? Análisis de la relación entre competitividad y desarrollo humano en México. *Revista del CLAD Reforma y Democracia* (51):1–20
- Otsuka K, Ali M (2020) Strategy for the development of agro-based clusters. *World Dev Persp*. <https://doi.org/10.1016/j.wdp.2020.100257>
- Porter M (1991) La ventaja competitiva de las naciones. Javier Vergara, Buenos Aires, Argentina
- Porter M (1999) Ser competitivo. Bilbao: Ediciones Deusto
- Putri EP, Chetchotsak D, Ruangchoenghum P, Jani MA, Hastijanti R (2016) Performance evaluation of large and medium scale manufacturing industry clusters in east java province, Indonesia. *Int J Technol* 7(7):1269–1279
- Quero L (2008) Estrategias competitivas: factor clave de desarrollo. *NEGOTIUM Revista Científica Electrónica Ciencias Gerenciales* 10(4):36–49
- Quiroga B, Lira L (2009) Técnicas de análisis regional. Series Manuales (59)
- Rabellotti R, Pietrobelli C (2005) Mejora de la competitividad en clusters y cadenas productivas en América Latina, el papel de las políticas. Washinton, D.C.: Banco Interamericano de Desarrollo
- Romanova A, Abdurakhmanov A, Ilyin V, Vygnanova M, Skrebutene E (2018) Formation of a regional industrial cluster on the basis of coordination of business entities' interests. *ICTE in Transportation and Logistics*
- Ruiz-Velazco A (2015) Territorial competitiveness and urban economic potential. *Rev Lider* 26:39–59
- Saavedra M. (2012). Una propuesta para la determinación de la competitividad en la pyme latinoamericana. *Pensamiento y gestión* (33):93–124
- Saavedra M, Milla S (2012) La competitividad en el nivel micro de la mipyme en el estado de Querétaro. XVII Congreso Internacional de
- Sarmiento Y (2019) Nociones generales del estudio de la competitividad territorial para planificar el desarrollo. *Retos De La Dirección* 13(1):103–116

- Sarmiento Y, González I, Pérez Y (2015) La competitividad territorial como insumo para la planificación. *Revista Cubana De Ciencias Económicas-EKOTEMAS* 1(3):1–17
- Sforzi F (1987) Identificazione spaziale. En G. Becattini, *Mercato e forze locali: il distretto industriale*, pp 143–167. il Mulino, Bologna
- Shakib MD (2020) Using system dynamics to evaluate policies for industrial clusters development. *Comput Ind Eng* 147:15. <https://doi.org/10.1016/j.cie.2020.106637>
- SIAP (s.f.) Recuperado el 30 de 09 de 2020, de [http://www.campomexicano.gob.mx/raw\\_pesca\\_gobmx/seccionar.php](http://www.campomexicano.gob.mx/raw_pesca_gobmx/seccionar.php)
- Soler V (2000) Verificación de las hipótesis del distrito industrial: una aplicación al caso valenciano. *Economía industrial* 4(334):13–23
- Suárez M (2005) La inserción de la pequeña y mediana empresa en el comercio exterior mexicano: un modelo de competitividad sistémica. Tesis de grado, UNAM, Facultad de Ciencias Políticas y Sociales
- UNESCO (s.f.) Research and development spending. Recuperado el 18 de 06 de 2019, de [http://uis.unesco.org/sites/all/modules/custom/uis\\_applications/apps/visualisations/research-and-development-spending/](http://uis.unesco.org/sites/all/modules/custom/uis_applications/apps/visualisations/research-and-development-spending/)
- Valencia K, Zetina A (2017) La cebolla mexicana: un análisis de competitividad en el mercado estadounidense, 2002–2013. *Región y sociedad* 29(70)
- Vázquez A, Reyes A (2013) Fundamentos sobre la competitividad para el desarrollo en el sector primario. *TLATEMOANI Rev Académica Investig* (14):1–29
- Wiratmadja II, Govindaraju R, Handayani D (2016) Innovation and productivity in indonesian it clusters: the influence of external economies and joint action. *Int J Technol* 7(6):1097–1106
- World Bank Group (s.f.) Obtenido de <https://datos.bancomundial.org/indicador/NE.GDI.FTOT.KD>
- World Bank Group (s.f.) Exportaciones de productos de alta tecnología (% de las exportaciones de productos manufacturados). Recuperado el 18 de 06 de 2019, de <https://datos.bancomundial.org/indicador/TX.VAL.TECH.MF.ZS?locations=MX-JP-KR-US&view=chart>
- World Bank Group (s.f.) Gasto en investigación y desarrollo (% del PIB). Recuperado el 18 de 06 de 2019, de <https://datos.bancomundial.org/indicador/GB.XPD.RSDV.GD.ZS?view=chart>
- World Bank Group (s.f.) Investigadores dedicados a investigación y desarrollo (por cada millón de personas). Recuperado el 18 de 06 de 2019, de <https://datos.bancomundial.org/indicador/SP.POP.SCIE.RD.P6>
- World Bank Group (s.f.) Solicitudes de patentes, residentes. Recuperado el 05 de 08 de 2019, de <https://datos.bancomundial.org/indicador/IP.PAT.RESD>
- World Economic Forum (2010) *The global competitiveness report 2010–2011*. Suiza, Geneva

# A Path Analysis of Learning Approaches, Personality Types and Self-Efficacy



Mine Aydemir and Nuran Bayram Arlı

**Abstract** The aim of this study is to explore how self-efficacy and personality types of undergraduates affect their learning approach and to analyze the relationship between the variables involved. A model was developed using self-efficacy, personality types and learning approach and this model was tested using path analysis. The path analysis showed that extraversion, neuroticism, conscientiousness and openness had a significant effect on self-efficacy, while extroversion and openness had a significant effect on both deep and surface learning. It was further found that self-efficacy had a significant effect on deep and surface learning. According to the results, personality types directly or/and indirectly affect the learning approaches. In light of the findings of this study, when the deep learning approach is considered as the desired learning approach, it can be said that the effects of self-efficacy and personality types on deep learning were remarkable.

**Keywords** Path analysis · Learning approaches · Self-efficacy · Personality types

## 1 Introduction

Human beings start to learn from the moment he/she is born and then continues to lifelong learning. Although learning processes are there in every moment of our lives, they are much more intense during our formal education. Students learn from predefined curricula throughout their education and they are taught using similar educational tools by educators. However, it is known that they do not perform at the same level and they may have problems learning specific concepts/subjects. Inability to learn a subject or a concept or inability to solve a problem is a result of several reasons. Among these are, the student, the teacher, the classroom, materials

---

M. Aydemir (✉) · N. Bayram Arlı  
Faculty of Economics and Administrative Sciences, Bursa Uludağ University, Bursa, Turkey  
e-mail: [mineaydemir@uludag.edu.tr](mailto:mineaydemir@uludag.edu.tr)

N. Bayram Arlı  
e-mail: [nuranb@uludag.edu.tr](mailto:nuranb@uludag.edu.tr)

used in the classroom, the learning environment, the method used for learning, the methods used for assessment, workload, etc. Each one of these factors have an effect on the learning process, and particularly important is the student himself/herself (Dart and Clarke 1991; Beattie et al. 1997; Biggs et al. 2001; Evans et al. 2003; Abraham et al. 2008). Even when the necessary conditions are satisfied for students' in terms of there may still be hindrances on their learning process or the learning may be slower than optimal. At this point, it is thought that other factors (self-esteem, their learning approach, their interests, their personality, etc.) may have an effect on learning process.

Learning approaches are an important issue in students' learning process. The learning approach adopted by the student has an effect on the learning outcomes throughout their studies. In this context, it is important to define and measure learning approaches. Among the reasons why it is wanted to define and measure the learning approaches of students are allowing them to be better learners, allowing the educators to perfect their teaching methods, defining problems in learning due to ineffective learning strategies and observation of the learning outcomes (Abraham et al. 2008; Dart and Clarke 1991).

## ***1.1 Theoretical Background***

Literature has different approaches to learning to offer. One of them is the Student Approaches to Learning (SAL) theory. SAL was developed by Marton and Säljö (1976). The authors coined the concept of learning approaches into two groups as deep and surface learning approaches. Deep and surface concepts were first explored by Craik and Lockhart (1972). In surface learning, the student pays minimum attention to the subject. She/he is interested in reproducing rather than understanding. She/he adopts an approach focusing on memorizing the content and experiences anxiety about their education. She/he does not show interest in the details of the subject and does not question them in depth. The student's motivation in surface learning is more extrinsic. According to Phan (2007), there is a positive relationship between habitual action and surface approach. On the other side, there is a positive relationship between learning complex processes and deep learning (Beattie et al. 1997; Phan 2011; Duff et al. 2004). The deep learning approach is about meaning and comprehension. In deep learning, the student communicates their questions about the subject, adopts a critical attitude and builds connections between new information and old ones. Students in the deep learning approach have intrinsic motivation. Deep learning approach is affected by factors such as personal experience and internal locus of control, while surface learning is affected by contextual factors (Biggs 1987a; Aharony 2006; Batı et al. 2010; Sæle et al. 2017).

### 1.1.1 Self-Efficacy and Learning Approaches

There are a number of studies emphasizing the importance of self-efficacy in overall learning processes. Self-efficacy, as the central element of the social cognitive perspective, is defined as a person's belief in the ability to perform a task or succeed in a situation. In this context, self-efficacy is about an individual's ability to produce results and their ability to perform at a specific level (Bandura 1993, 2006; Pajares and Valiante 1997; Phan 2011). Moreover, a person's ideas and their emotional reactions are influenced by their self-efficacy. People with higher self-efficacy perform better in difficult tasks and activities. Individual with lower self-efficacy, on the other hand, adopt a narrower perspective and think that the tasks or activities are harder than actually, they are. Self-efficacy and learning approaches are two theoretical frameworks in accounting for the learning and academic success of students (Bandura 1997; Pajares and Valiante 1997; Phan 2011). Learning approach and self-efficacy were analyzed for students and reported in the literature. A positive relationship is expected between self-efficacy and deep learning (Sins et al. 2008; Fenollar et al. 2007; Liem et al. 2008; Phan 2011; Çuhadar et al. 2013; Ekinçi 2015).

### 1.1.2 Learning Approaches and Personality Types

Learning and personality types relationship is not new. Different studies have been argued their relationship (Bidjerano and Dai 2007). Messick (1984) suggested that a person's learning style and their personality tendencies are consistent. Busato et al. (1999, 2000) explored learning approaches and personality types. It was found that meaning directed, i.e. deep learning had a positive relationship with openness/intellect. On the other hand, reproduction, i.e. surface learning was found to have a positive and direct relationship with compatibility and responsibility, while a positive and indirect relationship was found between neuroticism and reproduction. It is supposed that deep learning has a positive relationship with extraversion and openness/intellect (Costa and McCrae 1992), while surface learning has a positive relationship with neuroticism. Other studies reported a significant relationship between individual learning tendency and personality (Busato et al. 1999, 2000; Duff et al. 2004). According to the results of Chamorro-Premuzic and Furnham's (2009) study on first-year psychology students, the relation between learning approaches and personality types was lower than previously propounded. Also, according to Chamorro-Premuzic and Furnham (2009) in many studies only reported bivariate correlations and this is the lack of personality types and learning approaches studies.

### 1.1.3 Self-Efficacy and Personality Types

The concept of self-efficacy includes elements such as the planning of actions, organization of the necessary skills and the level of motivation as a result of reviewing the gains to be achieved with difficulties. Strong self-efficacy leads to success and



well-being (Bandura 1993; Yıldırım and İlhan 2010). Self-efficacy does not refer to being capable, but trusting one’s own resources. A person with sufficient skills to cope with a situation, but with low self-efficacy, will not be able to mobilize those skills. Self-efficacy is a general social cognition factor. We argue that personality and learning approach relations might work through self-efficacy. Research examining the correlation between personality types and self-efficacy, they found associations between them and some studies’ results showed that self-efficacy can be a mediator of personality types (Nauta 2004; Vecchione and Caprara 2009; Ebstrup et al. 2011; Löckenhoff et al. 2011; De Feyter et al. 2012; Stajkovic et al. 2018).

### 1.2 Aim and Hypotheses

This study was planned to answer these questions: Are the students’ personality types and self-efficacy beliefs effective in learning approaches? And these variables how important are for learning? To better understand the relationships between concepts and to find better solutions to problems, it is necessary to conduct studies focusing on this field and using different populations. Based on this information, the aim of this study was to model personality types and self-efficacy, factors believed to affect the learning approach of undergraduates and to describe their direct and indirect effects on the learning approach using path analysis.

In the light of this knowledge the proposed model is depicted in Fig. 1.

In accordance with previous research, we hypothesized the relationship between personality types, deep and surface learning and self-efficacy. In detail; positive effects of extraversion on deep learning (H<sub>1E</sub>) and surface learning (H<sub>2E</sub>), negative effects of neuroticism on deep learning (H<sub>3N</sub>) and positive effects on surface learning (H<sub>4N</sub>), positive effects of conscientiousness on deep learning (H<sub>5C</sub>) and negative

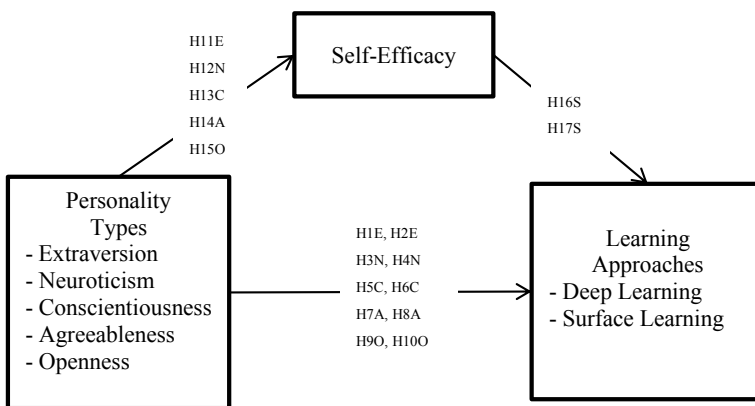


Fig. 1 A conceptual model

effects on surface learning ( $H_{6C}$ ), positive effects of agreeableness on deep learning ( $H_{7A}$ ) and negative effects on surface learning ( $H_{8A}$ ), positive effects of openness on deep learning ( $H_{9O}$ ) and negative effects on surface learning ( $H_{10O}$ ) were hypothesized. Positive effects of extraversion on self-efficacy ( $H_{11E}$ ), negative effects of neuroticism on self-efficacy ( $H_{12N}$ ), positive effects of conscientiousness on self-efficacy ( $H_{13C}$ ), negative effects of agreeableness on self-efficacy ( $H_{14A}$ ), positive effects of openness on self-efficacy ( $H_{15O}$ ) were hypothesized. Positive effects of self-efficacy on deep learning ( $H_{16S}$ ) and negative effects of self-efficacy on surface learning ( $H_{17S}$ ) were hypothesized. All hypotheses are based on the empirical results found in previous studies (Zhang 2003, Duff et al. 2004; Chamorro-Premuzic et al. 2007; Shokri et al. 2007; Chamorro-Premuzic and Furnham 2008; Sins et al. 2008; Lavasani et al. 2010; Phan 2011; Van Bragt et al. 2011; Wang et al. 2014; Ekinici 2015; Fosse et al. 2015; Wang et al. 2016).

## 2 Method

### 2.1 Participants

280 university students in the city of Bursa (Turkey) have participated as volunteers. A face to face questionnaire was done in order to collect data about their personality types, learning approaches, self-efficacy and demographic characteristics. The participants' ages ranged from 19 to 28 years, mean  $21.02 \pm 1.34$  (mean  $\pm$  sd) years. Hundred and ten (39%) were male and one hundred and seventy (61%) were female.

### 2.2 Measures

A questionnaire was submitted to the participants for data collection purposes and information on self-efficacy, learning approaches and personality types along with demographic data were collected. The convenience sampling method was preferred to collect data in this study. Among the scales used are The Revised Two Factor Study Process Questionnaire (R-SPQ-2F) by Biggs et al. (2001), General Self-Efficacy Scale (GSE) by Sherer et al. (1982) and Big Five Inventory (BFI) by John et al. (1991), John and Srivastava (1999). Turkish validity and reliability studies of all scales were previously conducted by different researchers.

The first scale, namely, Study Process Questionnaire (SPQ), was developed by Biggs (1987b) with the purpose of evaluation of SAL. Later, Biggs et al. (2001) developed a revised two-factor version of the Study Process Questionnaire (R-SPQ-2F). Several sample items are as follows: "I find that at times studying gives me a feeling of deep personal satisfaction.", "I do not find my course very interesting so I keep my work to the minimum.", "I work hard at my studies because I find the

material interesting.”, “I see no point in learning material which is not likely to be in the examination.” A 5-point Likert type scale was used. The Turkish version of this scale was studied by Batı et al. (2010). The scale consists of two sub-dimensions, namely deep learning approach and surface learning approach. And there are strategy and motivation groups under these sub-dimensions. Biggs et al. (2001) reported a Cronbach’s Alpha value of 0.74 for deep learning and 0.64 for surface learning. Batı et al. (2010), on the other hand, reported a Cronbach’s Alpha value of 0.77 for deep learning and 0.80 for surface learning. Consists of 20 items, the scale dedicates 10 questions for deep learning approach and 10 questions for surface learning approach. Higher scores on this scale show that the relevant learning approach is more commonly used. In this study, Cronbach’s Alpha value was found to be 0.75 for deep learning and 0.70 for surface learning.

The second scale used in this study was the GSE scale developed by Sherer et al. (1982) and Sherer and Adams (1983). Sample items include “When I make plans, I am certain I can make them work”, “I avoid facing difficulties”, “When unexpected problems occur, I don’t handle them well”, “I am a self-reliant person.” The original scale reported a Cronbach’s Alpha value of 0.86. The validity of the Turkish version of this scale was reported by Yıldırım and İlhan (2010). GSE consists of 17 items scales with a 5-point Likert type response scale. The scale’s Cronbach’s Alpha was found to be 0.80. Higher scores indicate higher self-efficacy. The Cronbach’s Alpha value found in this study for this scale was 0.87.

Finally, the Big Five Inventory scale was used in this study. BFI was developed by John et al. (1991), John and Srivastava (1999). Several sample items are as follows: “I am someone who is talkative”, “I am someone who can be somewhat careless”, “I am someone who tends to be lazy”, “I am someone who gets nervous easily”, “I am someone who is sophisticated in art, music, or literature”. A 5-point Likert type scale was used. The reliability of the Turkish version of this scale was reported by Alkan (2006). Cronbach’s Alpha ranged between 0.67 and 0.89. This scale consists of a total number of 44 items. Among the five personality types included in this scale are agreeableness, extroversion, neuroticism, conscientiousness and openness to experience. Higher scores on this scale are associated with the higher influence of the personality type in question (Ulu 2007). Cronbach’s Alpha values obtained in this study for five personality types were 0.77, 0.74, 0.73, 0.66 and 0.79 for extroversion, neuroticism, conscientiousness, agreeableness and openness to experience.

Path analysis was performed in this study to explore the relationships between learning approach, five-factor personality and self-efficacy. Analysis of Moment Structures (AMOS)-16 was used for path analysis and Statistical Package of Social Science (SPSS)-21 was used for all the other analyses.

### **2.3 Analysis**

Regression analysis is a standard analysis in modeling the relationships between the dependent variable and the independent variables. Path analysis is an analysis that

**Table 1** Model fit indices (Bayram 2016; Schumacker and Lomax 2004)

Model fit index	Suggested value	Acceptable value
$\chi^2/df$	$0 \leq \chi^2/df \leq 2$	$2 \leq \chi^2/df \leq 3$
RMSEA	$0 \leq RMSEA \leq 0.05$	$0.05 < RMSEA \leq 0.08$
SRMR	$0 \leq SRMR \leq 0.05$	$0.05 < SRMR \leq 0.10$
NFI	$0.95 \leq NFI \leq 1.00$	$0.90 \leq NFI < 0.95$
NNFI	$0.97 \leq NNFI \leq 1.00$	$0.95 \leq NNFI < 0.97$
CFI	$0.97 \leq CFI \leq 1.00$	$0.95 \leq CFI < 0.97$
GFI	$0.95 \leq GFI \leq 1.00$	$0.90 \leq GFI < 0.95$
AGFI	$0.90 \leq AGFI \leq 1.00$	$0.85 \leq AGFI < 0.90$

allows simultaneous modeling of relevant regression relationships. In path analysis, while a variable is a dependent variable in one model, it can be an independent variable in another model. Path analysis models are based on the work of Sewell Wright (1918, 1921, 1934, 1960). The drawing of such diagrams was first invented by him. Drawing path diagrams made it easier to understand, especially by visualizing simultaneous equation systems. Besides visualization, direct, indirect and total effects are also shown. In the 1960s and 1970s, many applications of path analysis were made in many sociology publications. Then, the applications spread to other social sciences areas. Using path analysis, complex relationships between variables can be modeled. The coefficients calculated by path analysis are standardized regression coefficients and indexes are used to evaluate model fit (Schumacker and Lomax 2010). Commonly reported fit indices can be listed as follows; Model Chi-Square  $\chi^2$ , (Adjusted) Goodness of Fit (AGFI/GFI), (Non) Normed Fit Index (NNFI/NFI), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA) and (Standardized) Root Mean Square Residual (SRMR/RMR). Model fit indices depict in Table 1.

Using path analysis the whole model was predicted together and each of the path coefficients was examined individually.

### 3 Findings

All scales used in the study (learning approaches, self-efficacy and big five personality) have high scores of reliability and they have been previously validated (Sherer et al. 1982; John et al. 1991; John and Srivastava 1999; Biggs et al. 2001; Ulu 2007; Bati et al. 2010; Yıldırım and İlhan 2010). The following Table 2 depicts the values and reliability coefficients found for the scales used and their sub-dimensions.

Alpha coefficients were calculated and all Cronbach’s Alpha coefficients found between 0.66 and 0.87 in all calculations involved. Table 3 shows the calculated correlation coefficients to evaluate the relationship between variables.

**Table 2** Descriptive statistics

Scales	Subscales	Items	Mean	Std.Dev	C. Alpha
Self-efficacy	Self-efficacy	17	63.49	9.56	0.87
Learning approach	Deep Learning	10	32.07	5.30	0.74
	Surface Learning	10	29.14	5.47	0.69
Big five personality	Extraversion	8	26.84	5.07	0.77
	Agreeableness	9	33.85	4.75	0.66
	Conscientiousness	9	33.02	5.25	0.73
	Neuroticism	8	24.39	5.27	0.74
	Openness	10	35.71	5.86	0.79

**Table 3** Correlation coefficients

	1	2	3	4	5	6	7
1 Self-efficacy	–						
2 Deep learning	0.38*	–					
3 Surface learning	-0.19*	-0.42*	–				
4 Extraversion	0.42*	0.09	0.05	–			
5 Agreeableness	0.22*	0.19*	-0.18*	0.10	–		
6 Conscientiousness	0.58*	0.33*	-0.16*	0.32*	0.43*	–	
7 Neuroticism	-0.25*	0.00	0.06	-0.31*	-0.08	-0.15**	–
8 Openness	0.39*	0.41*	-0.17*	0.31*	0.10	0.27*	-0.02

\* p < 0.01; \*\* p < 0.05

There were significant relationships among some of the personality types and learning approaches and also with self-efficacy. All personality types correlated with self-efficacy. Except for extraversion and neuroticism, the personality types were correlated with deep and surface learning. The following figure shows the estimated path model.

In Fig. 2, the estimated path model was calculated using standardized coefficients. Different fit indices were reported to evaluate the model. The resulting model had statistically significant and acceptable fit indices ( $X^2/df = 1.157$ ;  $p = 0.318$ ;  $RMSEA = 0.02$ ;  $SRMR = 0.04$ ;  $GFI = 0.99$ ;  $AGFI = 0.96$ ;  $CFI = 0.99$ ;  $NFI = 0.97$ ). Self-efficacy has a positive effect on deep learning ( $\beta = 0.22$ ;  $p < 0.01$ ), while it has a negative effect on surface learning ( $\beta = -0.18$ ;  $p < 0.01$ ). It was found that neuroticism ( $\beta = -0.12$ ;  $p < 0.01$ ), conscientiousness ( $\beta = 0.45$ ;  $p < 0.01$ ), extraversion ( $\beta = 0.18$ ;  $p < 0.01$ ) and openness ( $\beta = 0.22$ ;  $p < 0.01$ ) are among the personality types affecting self-efficacy. It was further found that deep learning and surface learning are affected by extroversion ( $\beta = -0.16$ ;  $p < 0.01$ ;  $\beta = 0.18$ ;  $p < 0.01$ ) and openness ( $\beta = 0.33$ ;  $p < 0.01$ ;  $\beta = -0.14$ ;  $p < 0.05$ ). Conscientiousness ( $\beta = 0.16$ ;  $p < 0.01$ ) affected on deep learning and agreeableness ( $\beta = -0.12$ ;  $p <$

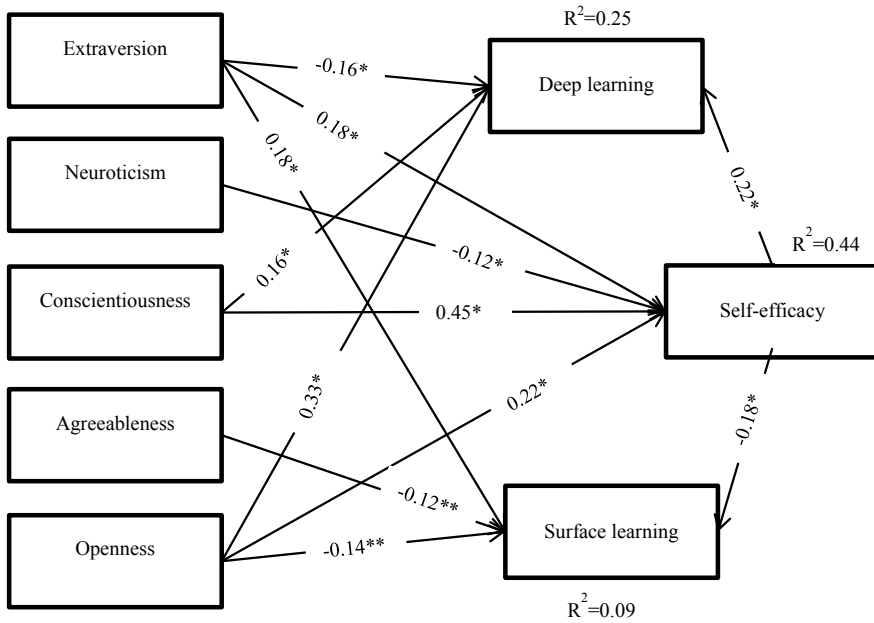


Fig. 2 Path model

0.05) affected on surface learning. Moreover,  $R^2$  value of self-efficacy was 0.44,  $R^2$  value of deep learning was 0.25, while it was 0.09 for surface learning. A number of fit indices were reported in this study and most of them were in the suggested range (Table 1).

Six hypotheses were not supported by the findings of the study. Negative effects of neuroticism on deep learning ( $H_{3N}$ ) and positive effects on surface learning ( $H_{4N}$ ), negative effects of conscientiousness on surface learning ( $H_{6C}$ ), positive effects of agreeableness on deep learning ( $H_{7A}$ ) and negative effects of agreeableness on self-efficacy ( $H_{14A}$ ). Also, the  $H_{1E}$  hypothesis was not supported because extraversion negatively affected deep learning.

## 4 Conclusion

The main contribution of our study is that it demonstrated that how self-efficacy and personality affected learning approaches in undergraduates students. In this study, the effect of the self-efficacy and personality types on learning approaches of undergraduates was modeled and estimated. Building on the literature reports, a theoretical model was built, having defined the relationship between these concepts and the direction of such relationships. The estimated model indicated that self-efficacy has a statistically significant effect on the learning approach. There was a positive

**Table 4** Conclusions with respect to the hypotheses

Hypothesis		Result
H <sub>1E</sub>	Extraversion → Deep learning	Rejected
H <sub>2E</sub>	Extraversion → Surface learning	Supported
H <sub>3N</sub>	Neuroticism → Deep learning	Rejected
H <sub>4N</sub>	Neuroticism → Surface learning	Rejected
H <sub>5C</sub>	Conscientiousness → Deep learning	Supported
H <sub>6C</sub>	Conscientiousness → Surface learning	Rejected
H <sub>7A</sub>	Agreeableness → Deep learning	Rejected
H <sub>8A</sub>	Agreeableness → Surface learning	Supported
H <sub>9O</sub>	Openness → Deep learning	Supported
H <sub>10O</sub>	Openness → Surface learning	Supported
H <sub>11E</sub>	Extraversion → Self-efficacy	Supported
H <sub>12N</sub>	Neuroticism → Self-efficacy	Supported
H <sub>13C</sub>	Conscientiousness → Self-efficacy	Supported
H <sub>14A</sub>	Agreeableness → Self-efficacy	Rejected
H <sub>15O</sub>	Openness → Self-efficacy	Supported
H <sub>16S</sub>	Self-efficacy → Deep learning	Supported
H <sub>17S</sub>	Self-efficacy → Surface learning	Supported

effect on deep learning and a negative effect on surface learning. The results of this study were in agreement with the expectations (Bandura 1997; Fenollar et al. 2007; Lavasani et al. 2010; Liem et al. 2008; Sins et al. 2008; Phan 2011; Ekinci 2015).

A closer look into how personality types affect learning approach showed that extroversion and openness to experience had a statistically significant effect on both learning approaches, conscientiousness has a direct effect on deep learning and agreeableness has a direct effect on surface learning. We did not find a relationship between neuroticism and learning approaches. Among the other studies which reported results similar to this study include (Busato et al. 1999, 2000; Chamorro-Premuzic et al. 2007; Chamorro-Premuzic and Furnham, 2008; Chamorro-Premuzic and Furnham, 2009; Duff et al. 2004). Previous studies have found a positive relationship between extraversion and deep learning. In these studies, there was no relationship between surface learning and extraversion (Zhang 2003, Duff et al. 2004, Chamorro-Premuzic et al. 2007, Chamorro-Premuzic and Furnham 2008). Differently, we found a negative relationship between extraversion and deep learning and a positive relationship between extraversion and surface learning.

Previous studies have found a positive relationship between agreeableness and deep learning (Chamorro-Premuzic et al. 2007; Shokri et al. 2007) and a negative relationship between agreeableness and surface learning (Zhang 2003, Chamorro-Premuzic et al. 2007). Whereas we found no relationship between agreeableness and deep learning. We found only negative relationship between agreeableness and surface learning.

In our findings, extraversion and openness to experience had both direct and indirect effects on learning approaches. Except the neuroticism, the other two personality types (conscientiousness and agreeableness) directly affected only one learning approach. As an important finding, agreeableness did not have a direct effect on self-efficacy. Other important findings, the neuroticism did not have a direct effect on learning approaches but has indirect effects. At this point, self-efficacy emerged as an important variable for the investigation of neuroticism personality type.

When we look at the relationships between personality types and self-efficacy, there were different studies that support our study. Authors reported that some personality types had a significant effect on self-efficacy (Wang et al. 2014, 2016; Brown and Cinamon 2016; Ebstrup et al. 2011; Skorek et al. 2014; Fosse et al. 2015). This study demonstrated that openness, extraversion and conscientiousness were all positively related to self-efficacy whereas neuroticism was negatively related to it. The agreeableness type did not have a statistically significant effect on self-efficacy.

Personality types and self-efficacy together explain 25% variance in deep learning and this shows that personality-related constructs have an important effect on deep learning in university students.

When all personality types, self-efficacy and learning approaches are simultaneously considered, all personality types except agreeableness have an effect on self-efficacy and self-efficacy relates to both learning approaches. It was seen that personality types and self-efficacy can affect learning approaches. These results showed that self-efficacy is an important social cognition factor when considering the relationship between personality and learning approaches.

We contribute to the understanding of influences of the personality types and self-efficacy on deep and surface learning. To our knowledge, it is the first time a relationship between personality and learning approaches has been tested with self-efficacy.

In light of the findings of this study, the deep learning approach is considered as the desired learning approach, it would be fair to say that the effect of self-efficacy and personality types on deep learning were remarkable. The existence of these and similar relations may be guiding in the organization of the educational system and in solving the problems related to the students. All of these characteristics are learnable. Students can learn strategies and can be more effective with their preferred style. In addition, understanding learning style differences can help teachers effectively reach most of the students.

The limitations of this study are the convenience sampling method used, the population focusing only on undergraduates and the use of cross-sectional data. The evaluation was based on self-report. Thus, future research is recommended to examine the personality types and moods of younger students in order to be better directed in the next years of their lives while gaining even more self-awareness. Future research should also look into other mechanisms that connect personality and self-efficacy.



## References

- Abraham RR, Vinod P, Kamath MG, Asha K, Ramnarayan K (2008) Learning approaches of undergraduate medical students to physiology in a non-PBL-and partially PBL-oriented curriculum. *Adv Physiol Educ* 32(1):35–37. <https://doi.org/10.1152/advan.00063.2007>
- Aharony N (2006) The use of deep and surface learning strategies among students learning english as a foreign language in an Internet environment. *Br J Educ Psychol* 76(4):851–866. <https://doi.org/10.1348/000709905X79158>
- Alkan N (2006) Reliability and validity of the Turkish version of the Big Five Inventory. Unpublished manuscript. University of Atılım, Ankara
- Bandura A (1993) Perceived self-efficacy in cognitive development and functioning. *Edu Psychol* 28(2):117–148. [https://doi.org/10.1207/s15326985Sep2802\\_3](https://doi.org/10.1207/s15326985Sep2802_3)
- Bandura A (1997) Self-efficacy: the exercise of control. W.H. Freeman, New York
- Batı AH, Tetik C, Gürpınar E (2010) Öğrenme yaklaşımları ölçeği yeni şeklini Türkçe'ye uyarlama ve geçerlilik güvenirlik çalışması. *Türkiye Klinikleri J Med Sci* 30(5):1639–1646. <https://doi.org/10.5336/medsci.2009-15368>
- Bayram N (2016) Yapısal Eşitlik Modellemesine Giriş AMOS Uygulamaları. 3b, Ezgi Kitabevi, Bursa
- Beattie V IV, Collins B, McInnes B (1997) Deep and surface learning: a simple or simplistic dichotomy? *Acc Educ* 6(1):1–12. <https://doi.org/10.1080/096392897331587>
- Bigderano T, Dai DY (2007) The relationship between the big-five model of personality and self-regulated learning strategies. *Learn Individ Differ* 17(1):69–81. <https://doi.org/10.1016/j.lindif.2007.02.001>
- Biggs J, Kember D, Leung DYP (2001) The revised two-factor study process questionnaire: R-SPQ-2F. *Br J Educ Psychol* 71:133–149. <https://doi.org/10.1348/000709901158433>
- Biggs JB (1987a) Study process questionnaire manual. Student approaches to learning and studying. Australian Council for Educational Research Ltd., Radford House, Frederick St., Hawthorn 3122, Australia. <https://files.eric.ed.gov/fulltext/ED308200.pdf>
- Biggs JB (1987b) Student approaches to learning and studying. Research monograph. Australian Council for Educational Research Ltd., Radford House, Frederick St., Hawthorn 3122, Australia. <https://files.eric.ed.gov/fulltext/ED308201.pdf>
- Brown D, Cinamon RG (2016) Personality traits' effects on self-efficacy and outcome expectations for high school major choice. *Int J Educ Vocat Guid* 16(3):343–361. <https://doi.org/10.1007/s10775-015-9316-4>
- Busato VV, Prins FJ, Elshout JJ, Hamaker C (1999) The relation between learning styles, the Big Five personality traits and achievement motivation in higher education. *Pers Individ Differ* 26(1):129–140. [https://doi.org/10.1016/S0191-8869\(98\)00112-3](https://doi.org/10.1016/S0191-8869(98)00112-3)
- Busato VV, Prins FJ, Elshout JJ, Hamaker C (2000) Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Pers Individ Differ* 29(6):1057–1068. [https://doi.org/10.1016/S0191-8869\(99\)00253-6](https://doi.org/10.1016/S0191-8869(99)00253-6)
- Chamorro-Premuzic T, Furnham A (2009) Mainly openness: the relationship between the Big Five personality traits and learning approaches. *Learn Individ Differ* 19(4):524–529. <https://doi.org/10.1016/j.lindif.2009.06.004>
- Chamorro-Premuzic T, Furnham A, Lewis M (2007) Personality and approaches to learning predict preference for different teaching methods. *Learn Individ Differ* 17(3):241–250. <https://doi.org/10.1016/j.lindif.2006.12.001>
- Costa PT Jr, McCrae RR (1992) Four ways five factors are basic. *Pers Individ Differ* 13:861–865. [https://doi.org/10.1016/0191-8869\(92\)90236-I](https://doi.org/10.1016/0191-8869(92)90236-I)
- Çuhadar C, Gündüz Ş, Tanyeri T (2013) Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü Öğrencilerinin Ders Çalışma Yaklaşımları ve Akademik Öz-Yeterlik Algıları Arasındaki İlişkinin İncelenmesi. *Mersin Univ J Faculty Educ* 9(1):251–259. <http://dergipark.gov.tr/download/article-file/160880>

- Dart BC, Clarke JA (1991) Helping students become better learners: a case study in teacher education. *High Educ* 22(3):317–335. <https://doi.org/10.1007/BF00132294>
- De Feyter T, Caers R, Vigna C, Berings D (2012) Unraveling the impact of the Big Five personality traits on academic performance: the moderating and mediating effects of self-efficacy and academic motivation. *Learn Individ Differ* 22(4):439–448. <https://doi.org/10.1016/j.lindif.2012.03.013>
- Duff A, Boyle E, Dunleavy K, Ferguson J (2004) The relationship between personality, approach to learning and academic performance. *Pers Individ Differ* 36(8):1907–1920. <https://doi.org/10.1016/j.paid.2003.08.020>
- Ebstrup JF, Eplöv LF, Pisinger C, Jørgensen T (2011) Association between the Five Factor personality traits and perceived stress: is the effect mediated by general self-efficacy?. *Anxiety Stress Coping* 24(4):407–419. <https://doi.org/10.1080/10615806.2010.540012>
- Ekinçi N (2015) The relationships between approaches to learning and self-efficacy beliefs of candidate teachers. *Hacettepe Univ J Educ* 30(1):62–76. <http://www.efdergi.hacettepe.edu.tr/volume-30-issue-1-year-2015.html>
- Evans CJ, Kirby JR, Fabrigar LR (2003) Approaches to learning, need for cognition, and strategic flexibility among university students. *Br J Educ Psychol* 73(4):507–528. <https://doi.org/10.1348/000709903322591217>
- Fenollar P, Román S, Cuestas PJ (2007) University students' academic performance: an integrative conceptual framework and empirical analysis. *Br J Educ Psychol* 77(4):873–891. <https://doi.org/10.1348/000709907X189118>
- Fosse TH, Buch R, Säfvenbom R, Martinussen M (2015) The impact of personality and self-efficacy on academic and military performance: the mediating role of self-efficacy. *J Mil Stud* 6(1):47–65. <https://doi.org/10.1515/jms-2016-0197>
- John OP, Srivastava S (1999) The Big-Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin LA, John OP (eds) *Handbook of personality: theory and research*. Guilford Press, New York, pp 102–138. [http://moityca.com.br/pdfs/bigfive\\_john.pdf](http://moityca.com.br/pdfs/bigfive_john.pdf)
- John OP, Donahue EM, Kentle RL (1991) The big five inventory—versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA. [http://www.sjdm.org/dmidi/Big\\_Five\\_Inventory.html](http://www.sjdm.org/dmidi/Big_Five_Inventory.html)
- Lavasani MG, Malahmadi E, Amani J (2010) The role of self-efficacy, task value, and achievement goals in predicting learning approaches and mathematics achievement. *Proc Soc Behav Sci* 5:942–947. <https://doi.org/10.1016/j.sbspro.2010.07.214>
- Liem AD, Lau S, Nie Y (2008) The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemp Educ Psychol* 33:486–512. <https://doi.org/10.1016/j.cedpsych.2007.08.001>
- Löckenhoff CE, Duberstein PR, Friedman B, Costa PT Jr (2011) Five-factor personality traits and subjective health among caregivers: the role of caregiver strain and self-efficacy. *Psychol Aging* 26(3):592. <https://doi.org/10.1037/a0022209>
- Nauta MM (2004) Self-efficacy as a mediator of the relationships between personality factors and career interests. *J Career Assess* 12(4):381–394. <https://doi.org/10.1177/1069072704266653>
- Pajares F, Valiante G (1997) Influence of self-efficacy on elementary students' writing. *J Educ Res* 90(6):353–360. <https://doi.org/10.1080/00220671.1997.10544593>
- Phan HP (2011) Interrelations between self-efficacy and learning approaches: a developmental approach. *Educ Psychol* 31(2):225–246. <https://doi.org/10.1080/01443410.2010.545050>
- Sæle RG, Dahl TL, Sørli T, Friberg O (2017) Relationships between learning approach, procrastination and academic achievement amongst first-year university students. *High Educ* 74(5):757–774. <https://doi.org/10.1037/0022-0663.84.3.261>
- Schumacker RE, Lomax R (2004) *A beginner's guide to SEM*, 2nd edn. Lawrence Erlbaum Associates Publishers, New Jersey
- Schumacker RE, Lomax R (2010) *A beginner's guide to SEM*, 3rd edn. Taylor&Francis, Routledge
- Sherer M, Adams CH (1983) Construct validation of the self-efficacy scale. *Psychol Rep* 53(3):899–902. <https://doi.org/10.2466/pr0.1983.53.3.899>

- Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, Rogers RW (1982) The self-efficacy scale: construction and validation. *Psychol Rep* 51(2):663–671. <https://doi.org/10.2466/pr0.1982.51.2.663>
- Shokri O, Kadivar P, Farzad VE, Sangari AA (2007) Role of personality traits and learning approaches on academic achievement of university students. *Psychol Res* 9:65–84. <https://www.sid.ir/en/journal/ViewPaper.aspx?id=102080>
- Sins PHM, Van Joolingen WR, Savelsbergh ER, Van Hout-Wolters B (2008) Motivation and performance within a collaborative computer-based modeling task: relations between students' achievement goal orientation, self-efficacy, cognitive processing, and achievement. *Contemp Educ Psychol* 33:58–77. <https://doi.org/10.1016/j.cedpsych.2006.12.004>
- Skorek M, Song AV, Dunham Y (2014) Self-esteem as a mediator between personality traits and body esteem: path analyses across gender and race/ethnicity. *PLoS one* 9(11):e112086. <https://doi.org/10.1371/journal.pone.0112086>
- Stajkovic AD, Bandura A, Locke EA, Lee D, Sergent K (2018) Test of three conceptual models of influence of the big five personality traits and self-efficacy on academic performance: a meta-analytic path-analysis. *Pers Individ Differ* 120:238–245. <https://doi.org/10.1016/j.paid.2017.08.014>
- Ulu IP (2007) An investigation of adaptive and maladaptive dimensions of perfectionism in relation to adult attachment and big-five personality traits. Doctoral Dissertation, Middle East Technical University, Ankara. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Van Bragt CA, Bakx AW, Bergen TC, Croon MA (2011) Looking for students' personal characteristics predicting study outcome. *High Educ* 61(1):59–75. <https://doi.org/10.1007/s10734-010-9325-7>
- Vecchione M, Caprara GV (2009) Personality determinants of political participation: the contribution of traits and self-efficacy beliefs. *Pers Individ Differ* 46(4):487–492. <https://doi.org/10.1016/j.paid.2008.11.021>
- Wang Y, Yao L, Liu L, Yang X, Wu H, Wang J, Wang L (2014) The mediating role of self-efficacy in the relationship between Big five personality and depressive symptoms among Chinese unemployed population: a cross-sectional study. *BMC Psychiatry* 14(1):61. <https://doi.org/10.1186/1471-244X-14-61>
- Wang JH, Chang CC, Yao SN, Liang C (2016) The contribution of self-efficacy to the relationship between personality traits and entrepreneurial intention. *High Educ* 72(2):209–224. <https://doi.org/10.1007/s10734-015-9946-y>
- Yıldırım F, İlhan İÖ (2010) Validity and reliability study of the Turkish self-efficacy scale. *Turkish J Psychiatry* 21(4):301–308. <http://www.turkpsikiyatri.com/pdf/c21s4/301-308.pdf>
- Zhang LF (2003) Does the big five predict learning approaches? *Pers Individ Differ* 34(8):1431–1446

# Emotions Mining Research Framework: Higher Education in the Pandemic Context



Radka Nacheva

**Abstract** The pandemic situation in 2020 was a challenge for the organization of the educational process in higher education. The crisis has exacerbated inequalities between universities, which funding, digital sustainability and emergency training are weaker than their national and international competitors. The poor provision of the learning process in an electronic environment has led to a number of problems. Some of them are related to the acquisition of learning material and practical skills, lack of communication between students and academic staff. The increased use of the Internet during periods of social distance has also led to an increase in participating in social media activities, which have become forums for sharing opinions and expressing emotions through text and multimedia content. In this regard, the aim of the article is to propose a research framework for evaluation of emotional attitudes in social media. The author tested the practical applicability of the proposed framework by retrieving data from the social network Twitter and applying data mining techniques for analyzing large volumes of textual content.

**Keywords** Emotions mining · Text mining · Social networks · Higher education · Pandemic

## 1 Introduction

The pandemic, which began in 2020, has faced many challenges for humanity. Many sectors have been severely affected by lockdown in countries, leading to the closure of companies or the reduction of employment in a number of sectors. The consequences are a lessening in company revenues, rising unemployment, deteriorating credit ratings of citizens and businesses and even loss of property. The International Monetary Fund (IMF) reports a deepening recession in most countries in 2020—an average of 4.4% shrinking in the global economy (International Monetary Fund 2020). According to the World Bank, that is the worst reported since World War II

---

R. Nacheva (✉)  
University of Economics – Varna, 77, Knyaz Boris I Blvd., Varna, Bulgaria  
e-mail: [r.nacheva@ue-varna.bg](mailto:r.nacheva@ue-varna.bg)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_18](https://doi.org/10.1007/978-3-030-85254-2_18)

299

(World Bank 2020). For the European Union, the rate of decline is even higher – 7.6% by October 2020 (International Monetary Fund 2020).

The education sector has not remained unaffected by the crisis and has undergone a number of changes to ensure the learning process in schools and universities. Despite efforts at the state, institutional and personal levels, according to UNICEF 1.6 billion scholars and students from 188 countries were affected by the crisis in April 2020 due to the closure of schools and universities (UNICEF 2020). A decline of up to 1 billion is observed for the new academic year. That is why the processes of digitalization in educational institutions are being forced. Among the tools for providing education offered by some authors are the introduction of mobile learning (Todoranova and Penchev 2020), application of artificial intelligence technologies (Petrova and Sulova 2020), gamification software (Stoyanova 2015), social networks (Malkawi et al. 2021). Some researchers also report a number of barriers to digital learning, such as the high cost of software solutions, including their acquisition and maintenance (Kuyumdzhiiev 2020). This requires development of regional and national policies related to the improvement of technical, social and economic infrastructure to build conditions for the rise of educational institutions (Czaplewski and Klóska 2020).

According to the optimistic forecasts of the IMF, from the beginning of 2021 the average global growth is expected to be 5.2%, and for the EU 5% (International Monetary Fund 2020). Unfortunately, skeptical opinions are expressed regarding the recovery from the crisis. The reason is that the pandemic had a negative impact on the accumulation of human capital, and hence on active job seeking (World Bank Group 2021). Here it should be borne in mind that the quality of the workforce is an important factor in increasing the competitiveness of companies (Antonova and Ivanova 2018). This increases the requirements of the labor market, respectively to the educational institutions. The World Bank is developing scenarios for global economic growth and possible solutions to the crisis, which depend on the pace of pandemic control (World Bank Group 2021). One of the main emphases placed in them is investing funds in the development of human capital through the implementation of adequate policies in the field of education.

In this regard, **the aim of the article** is to propose a research framework for evaluation of emotional attitudes in social media. The purpose is achieved through the following objectives:

- Studying the higher education issues in a pandemic context;
- Investigating the approaches to social media mining which are applicable to emotions mining.

The author tested the practical applicability of the proposed framework by retrieving data from the social network Twitter and applying data mining techniques for analyzing large volumes of textual content.

## 2 Higher Education Challenges During Pandemic

In the conditions of intensified competition between universities, both nationally and internationally, one of the challenges facing them is competitive differentiation. Many of them fail to implement adequate marketing strategies to stakeholders. Successful branding must help to achieve a clear differentiation of directly competing universities through unique attributes and characteristics that are relevant to the target groups (Zhechev 2018). The rules of branding in higher education have changed with the onset of the Covid-19 crisis. It has seriously affected the attraction of new students. According to a study by the American Marketing Association from October 2020, 72% of university rectors (respectively presidents) are concerned about the decline in the value of higher education in the pandemic compared to March (48%) and April (60%) of the same year (American Marketing Association 2020). Unfortunately, there are also negative opinions among students—56% said that they can no longer afford to continue their education due to financial problems during the crisis. 36% of the parents interviewed said that they have redirected the funds for financing their children's education to cover expenses or financial losses incurred as a result of the Covid-19 crisis.

The problems that young people face during lockdown periods are far from limited to their deteriorating financial situation. A study by the University of Copenhagen shows that stress levels among people under the age of 30 increased during the pandemic period (Rohde 2021). It has been found that in participants without previous mental problems, there are those that deepen with the extension of the lockdown period. Social distance raises some physical and mental problems that lead to a decrease or lack of motivation to perform duties in a learning environment.

In an interview with the World Economic Forum, Prof. Suzanne Fortier, the Principal and Vice-Chancellor of McGill University in Montreal, Canada, a Chair of the Global University Leaders Forum too, outlined some problems in higher education that arose during periods of social distance (Fleming 2021). These include: lack of lecturer–student communication; difficulties in acquiring new knowledge and practical skills; difficulties in conducting classes, meetings, conversations due to various technical and financial reasons; problems with the social inclusion of people with special needs or people from other cultural communities. On the other hand, students in the above courses, those who come to the universities for upskilling and reskilling, typically people who are already in the workforce, have found many advantages in the flexibility of e-learning.

In addition, the challenges outlined by Bhagat and Kim can be highlighted (Bhagat and Kim 2020): lowering the quality of education as a result of distance learning; digital sustainability of universities; preparatory to and flexibility of the academic staff in providing educational services in an electronic environment; creating adequate policies for tuition, conducting exams and acquiring educational degrees that correspond correctly to the "new" reality.

Stanimirov's analysis of the educational environment states that "recognizing the challenges is the basis for generating ideas for 'closing the gaps' and synchronizing

the education and labour markets, which have a cycle of 3 to 5 years" (Stanimirov 2020). As the demands of society increase, so do the demands on universities, especially in conditions of social distance, when they have to show their digital resilience and adaptability.

Given the above, as well as in view of some statistics<sup>1</sup> for the increased use of the Internet during the periods of social distance in 2020, it can be outlined the following main tasks of this article:

- To suggest an approach to research the emotional attitudes of Internet users, in particular in social networks;
- To study the social networks users' opinions, mining emotional attitudes about periods of social distance.

In order to accomplish the tasks, it is necessary to study approaches to social media emotions mining.

### 3 Social Media Mining Approaches and Technologies

Undoubtedly, social media is a big data source. One of the most popular technologies for emotions mining in social media is based on extracting knowledge from data. Data extracted from social media are unstructured and often difficult to process due to their diverse nature. The extracted content can be textual or multimedia, which implies the application of various techniques for its processing.

Sulova and Bankov suggest a four-stage approach to social media mining: Data retrieval; Text processing; Data mining and Results interpretation (Sulova and Bankov 2019). They emphasize that their proposed approach can be adapted to the specifics of using a particular social media.

Manguri et. al. apply an approach to extracting feelings from social networks, which consists of the following stages: Text data mining; Sentiment identification; Feature selection; Sentiment classification; Sentiment polarity & subjectivity (Manguri et al. 2020). As a basis for testing the approach, the authors use textual content from the social network Twitter.

Nasralah et. al. follow a comprehensive approach to retrieving textual content from social networks, which is carried out at the following main stages (Nasralah et al. 2020): Discovery and topic detection, Data collection, Data preparation and quality evaluation and Analysis and results (Fig. 1).

Another popular technology for working with large amounts of unstructured data, that we find in the literature, is artificial neural network (ANN). According to Bakaev

---

<sup>1</sup> According to Eurostat, on average in the European Union, 26% of users have used the Internet to conduct e-learning activities, 56% for social networking activities, 68% for messaging, 74% for working with e-mail. The highest use of social networks was reported in Denmark 85%, and the lowest in Germany 54% (Eurostat 2021).



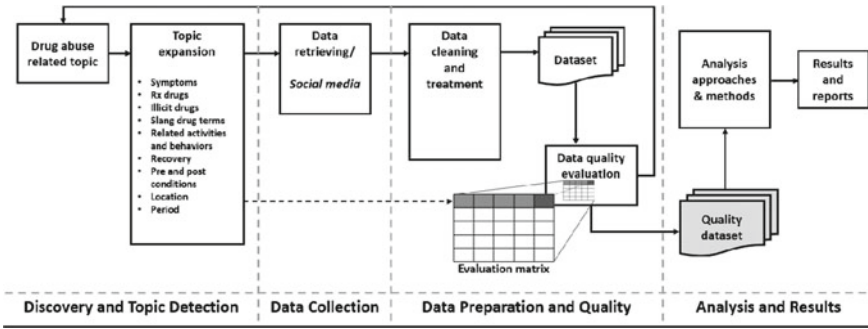


Fig. 1 Social media text mining framework by Nasralah et al. (2020)

et. al., it serves to construct a model of consumer behavior and predict the complexity of the Internet resource with which the user works (Bakaev et al. 2018).

Krebs et. al. follow a mixed approach to emotions mining in social networks based both on neural network technology and text mining (Krebs et al. 2018). Their approach, called "Pipeline for final prediction of reaction distributions", consists of several stages that take place simultaneously (Fig. 2).

Calefato et. al. propose a framework architecture of emotions mining from textual content. It consists of two main modules: Emotion Classification Module and Polarity Classification Module (Calefato et al. 2019). The authors define it as a specific solution for sentiment analysis, specifically for polarity and emotions mining from a text.

Yassine and Hajj suggest a framework architecture for emotions mining from textual content on social networks (Yassine and Hajj 2010). It consists of seven

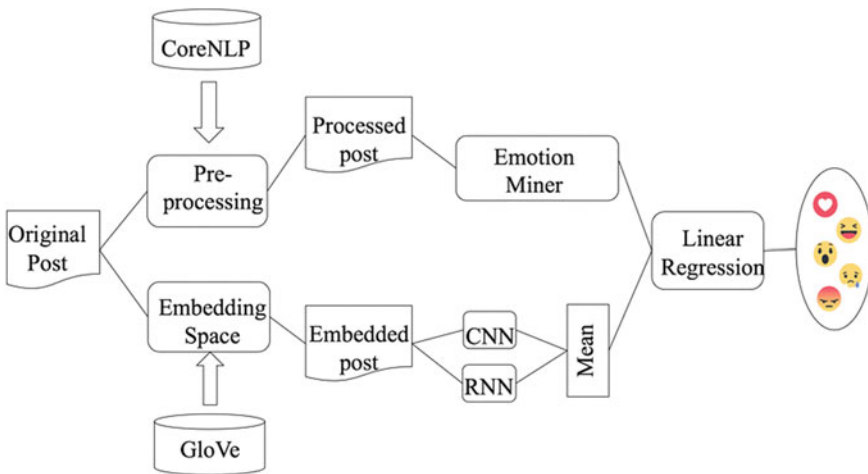


Fig. 2 Pipeline for final prediction of reaction distributions by Krebs et al. (2018)



steps: Raw data collection, Lexicon’s development, Feature generation, Data pre-processing, Creating a training model for text subjectivity, Text subjectivity classification, Friendship Classification.

Luo and Yi’s approach to emotions mining from online comments consist of: Dataset and Pre-processing, Setting Parameters and Determining the Number of Topics, Comparison of Sampling Time with Existing Models, Comparison of Sampling Time with Existing Models, Understandability of Results (Luo and Yi 2019).

On the basis of the cited publications, it can be concluded that the variety of approaches and technologies for emotions mining is huge. Scientists mainly focus on text mining and neural networks. That is why there are some shared stages in the approaches considered: Retrieving social media content; Content pre-processing; Classification of the content according to the identified emotions; Interpretation of results.

### 4 Emotions Mining Research Framework

In view of the purpose of this study and based on what has been stated so far, the author proposes a research approach to assessing emotional attitudes in social media (Fig. 3). It consists of three swim lanes: Stages, Toolkit and Artifacts (considered from bottom to top).

In this article, the author proposes an evaluation process of emotional attitudes in social media which is conducted in five stages: Topics Discovery, Social Media Connecting, Data Retrieving and Pre-processing, Modeling and Classification, Analysis and Evaluation.

For the implementation of the phases, tools for extracting, processing and evaluating content from social media are used. The output of each of the phases, resulted from a certain toolkit, are one or more artifacts: Research Plan, Data Source, Datasets, Models and Reports.

It should be noted that the proposed framework adapts to the specifics of the social media which content is being examined. For example, each social media application

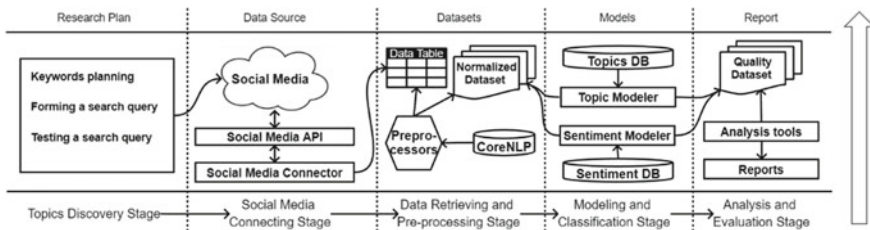


Fig. 3 Proposed Emotions Mining Research Framework

programming interface (API) has strictly individual characteristics and requires the use of different connectors for different social media mining platforms.

On Topics Discovery Stage tools for planning keywords and queries to social media are applied. This type of software evaluates the relevance of keywords and offers a set of words to form an effective query.

Based on the selected keywords set, a request is made to social media. It is necessary to use specialized data mining software. It connects to the social media API via a connector, after which the raw set of unstructured data is generated. The data mining software supports a set of pre-processors that use natural language processing methods (CoreNLP). Most often they are: Transformation, Tokenization, N-grams, Filtering, Normalization. They are applied in the order in which they are listed, and their settings depend on the objectives of the study.

As a result, a set of data is generated, which is used as a basis for the implementation of the next phase of modeling and classification. The topics are identified and the emotional attitudes are analyzed according to the classification databases. These databases contain a pre-prepared set of words related to a specific research field.

The last phase is related to the analysis of the generated models from the previous stage. Reports with the results of the study are created. Typically, they contain statistical analysis data and conclusions about the overall emotional attitudes in the specified research questions.

## 5 Results

It was used a software set to test the proposed framework. These are: Gephi, MeaningCloud and Orange. Gephi is an open-source and free visualization and exploration software for all kinds of graphs and networks. MeaningCloud is an Excel Add-in for text analytics (MeaningCloud 2021). Orange is an open-source data visualization, machine learning and data mining toolkit (Orange 2021).

The social network Twitter was used to test the proposed research framework. In the period 24.01.–26.01.2021 7722 tweets in English were retrieved using several keywords: university e-learning 2020; university COVID-19; financial crisis education 2020; education challenges; education crisis 2020; lockdown; higher education crisis; COVID-19 crisis 2020; education 2020; higher education COVID-19. The experiment is based on extracted text content.

Gephi was used to connect to Twitter and retrieve data. It is a powerful tool based on algorithms for visualizing and simulating graphs. It is characterized by speed. It provides a statistics and metrics framework (Gephi 2021). Downloaded tweets are exported as a csv file.

The next step is to process the raw data via Orange. The analysis of emotional attitudes is performed using the Tweet Profiler module (Fig. 4), which supports several methods of content classification. These are classes based on the classifications of Plutchnik, Ekman and Profile of Mood States (POMS). Each of them identifies a different number of basic emotions.



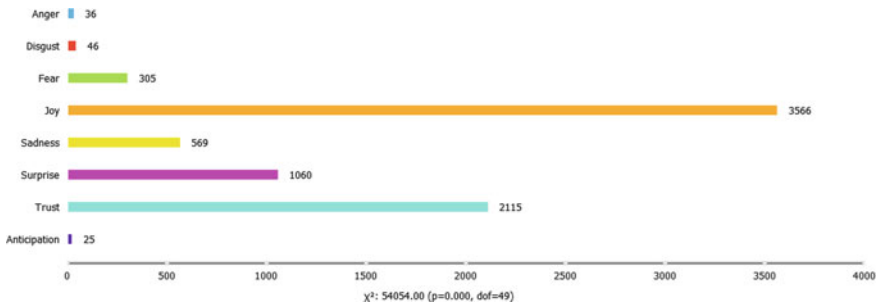
**Fig. 4** Configuration of emotion analysis in Orange

According to Plutchik, these are: anticipation, acceptance, joy, surprise, anger, disgust, fear, sadness (Plutchik 1980). Ekman distinguishes joy, surprise, anger, disgust, fear, sadness (Ekman 1982). POMS classifies the emotions of tension, anger, vigor, fatigue, depression, confusion (Renger 1993). The common emotion for the three classifications is anger, which is considered a primary negative emotion, thanks to which individuals defend and survive, both physically and verbally.

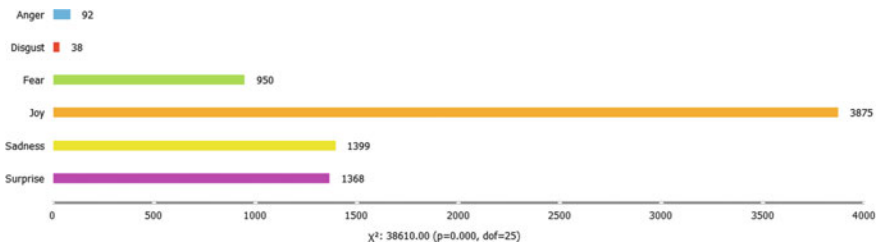
The results generated by Orange after applying the Plutchnik classifier are shown in Figs. 5. It is noticed that the positive emotions joy and trust are the highest percentage—73.56% of all extracted tweets.

The results according to the Ekman classifier (Fig. 6) are similar to the previous one. Positive emotion joy also gives precedence—50.18% of all tweets.

In Fig. 7 it is noticed that the diagram changes when applying POMS. The reason is that this classification is oriented entirely to negative emotions.



**Fig. 5** Results of the emotions analysis in Orange according to the classifier of Plutchnik



**Fig. 6** Results of the emotions analysis in Orange according to the classifier of Ekman

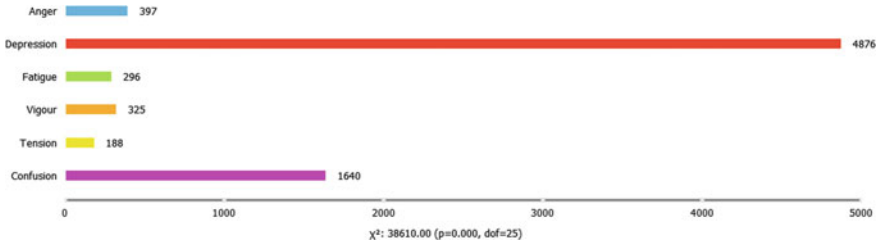


Fig. 7 Results of the emotions analysis in Orange according to the classifier of POMS

The differences in the results of the application of the different classifiers also arise from the pre-defined set of words that is applied for sentiment analysis of Orange.

For comparison, we apply MeaningCloud—an Excel plugin, through which we perform sentiment analysis. We apply the built-in basic model in the software based on WordNet. The results are summarized in Table 1. Similar to Plutchik’s and Ekman’s Orange classifiers, MeaningCloud recognizes that 55% of the content analyzed is positive polarity. 14% is marked with neutral polarity and 31% with negative polarity.

Therefore, it can be concluded that the emotional attitudes identified in the extracted tweets are predominantly positive. There are observed nuances of negative emotions, including anger, fear, disgust and depression.

According to the research framework proposed in this article, the next step is identifying the topics that excite the Twitter users. For this purpose, we use the Orange modules for Pre-processing and Topic Modeling (Fig. 8).

The applied pre-processors are:

Table 1 Relative share of polarity types

Polarity	Tweets Count	Relative share (%)
P +	1168	15
P	3077	40
NEU	1043	14
N	1469	19
N +	965	12

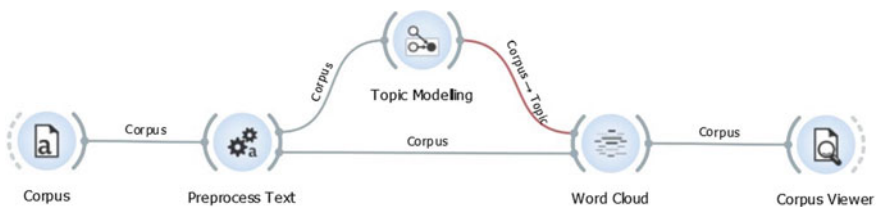


Fig. 8 Orange theme modeling configuration



mainly in a digital environment. Over time, the “new” reality was accepted by society and people adapted their daily lives to it.

These preconditions have led the author of this article to conduct a study of the emotional attitudes of social media users, in particular the social network Twitter, based on appropriate keywords. As a result, a framework has been proposed that can be adapted depending on the needs of the research and the specific features of social media.

**Acknowledgements** The publication is made within project No. 8.2.2.0/18/A/021 “Perfection of the Academic Staff of Liepaja University in the Areas of Strategic Specialization—Natural Sciences, Mathematics and Information Technologies, Art, Social Sciences, Commerce and Law”.

## References

- American Marketing Association (2020) 2020 Symposium for the marketing of higher education, November 16–19, [https://www.ama.org/wp-content/uploads/2020/11/2020-Higher-Ed\\_Resource-Guide\\_v4.pdf](https://www.ama.org/wp-content/uploads/2020/11/2020-Higher-Ed_Resource-Guide_v4.pdf). Accessed 27 Jan 2021
- Antonova K, Ivanova P (2018) Staff leasing. *Oxford Economic Papers*, Oxford University Press, vol 70, 4, issue 2, pp 1416–1425
- Bakaev M, Laricheva T, Heil S, Gaedke M (2018) Analysis and prediction of university websites perceptions by different user groups. In: *Proceedings of XIV international scientific-technical conference on actual problems of electronics instrument engineering (APEIE)*, IEEE, pp 381–385
- Bhagat S, Kim D (2020) Higher education amidst COVID-19: challenges and silver lining. *Inf Syst Manag* 37(4):366–371
- Calefato F, Lanubile F, Novielli N, Quaranta L (2019) EMTk—the emotion mining toolkit. In: *IEEE/ACM 4th international workshop on emotion awareness in software engineering (SEmotion)*. Montreal, QC, Canada, pp 34–37
- Czaplewski M, Klóska R (2020) Regional policy as a factor in shaping regional development in Poland. *South East European J Econ Bus* 15(1):93–104
- Ekman P (1982) *Emotion in the human face*, 2nd edn. Cambridge University Press, New York
- Eurostat (2021) What did we use the internet for in 2020? <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210126-2>. Accessed 27 Jan 2021
- Fleming S (2021) This is how university students can emerge from the pandemic stronger, <https://www.weforum.org/agenda/2021/01/online-learning-universities-covid-suzanne-fortier/>. Accessed 27 Jan 2021
- Gephi (2021) Gephi features, <https://gephi.org/features/>. Accessed 27 Jan 2021
- International Monetary Fund (IMF) (2020) Real GDP growth: annual percent change, [https://www.imf.org/external/datamapper/NGDP\\_RPCH@WEO/OEMDC/EU/ADVEC/ASS/DA/WEOWORLD](https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/EU/ADVEC/ASS/DA/WEOWORLD). Accessed 27 Jan 2021
- Krebs F, Lubascher B, Moers T, Schaap P, Spanakis G (2018) Social emotion mining techniques for facebook posts reaction prediction. In: *Proceedings of the 10th international conference on agents and artificial intelligence*, vol 1, pp 211–220
- Kuyumdzhiiev I (2020) A model for timely delivery of IT solutions for Bulgarian universities. In: *International multidisciplinary scientific geoconference: SGEM 20(2.1.)*, pp 3–10
- Luo X, Yi Y (2019) Topic-specific emotion mining model for online comments. *Future Internet* 11(e79):18
- Malkawi R, Alsmadi I, Ahmed A, Petrov P (2021) A firewall-adversarial testing approach for software defined networks. *J Theoret Appl Inform Technol*, Little Lion Scientific 99(1):227–241

- Manguri K, Ramadhan R, Mohammed Amin P (2020) Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan J Appl Res* 5(3):54–65
- MeaningCloud (2021) MeaningCloud official website, <https://www.meaningcloud.com/>. Accessed 27 Jan 2021
- Nasralah T, El-Gayar O, Wang Y (2020) Social media text mining framework for drug abuse: development and validation study with an opioid crisis case analysis. *J Med Internet Res* 22(8):e18350
- Orange (2021) Orange official website, <https://orangedatamining.com/>. Accessed 27 Jan 2021
- Petrova R, Sulova S (2020) AI Governor for the quality and the strength of bridges. *CompSysTech '20*. In: Proceedings of the 21st international conference on computer systems and technologies 20. Association for Computing Machinery, pp 78–85
- Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H (eds) *Emotion: theory, research, and experience*, vol 1, Academic Press Cambridge, pp 3–31
- Renger R (1993) A review of the profile of mood states (POMS) in the prediction of athletic success. *J Appl Sport Psychol* 5(1):78–84
- Rohde A (2021) Younger people are being hit hardest by pandemic loneliness, <https://www.weforum.org/agenda/2021/01/pandemic-loneliness-covid19-mental-health>. Accessed 27 Jan 2021
- Stanimirov E (2020) Quo Vadis, Education? In: Proceedings of Jubilee international scientific conference dedicated to the 100th anniversary of the university of economics—varna economic science, education and the real economy: development and interactions in the digital age, vol 1. Publishing house Science and Economics, pp 27–49 (in Bulgarian)
- Stoyanova M (2015) Architecture of gamification system. *Sci j Econ Comp Sci* 2:18–33 (in Bulgarian)
- Sulova S, Bankov B (2019) Approach for social media content-based analysis for vacation resorts. *J Commun Softw Syst (JCOMSS)* 15(3):262–271
- Todoranova L, Penchev B (2020) A conceptual framework for mobile learning development in higher education. In: Proceedings of the 21st international conference on computer systems and technologies '20 (CompSysTech '20). Association for Computing Machinery, pp 251–257
- UNICEF (2020) COVID-19 and children: UNICEF data hub, <https://data.unicef.org/covid-19-and-children/>. Accessed 27 Jan 2021
- World Bank Group (2021) Global economic prospects. World Bank, Washington
- World Bank (2020) COVID-19 to plunge global economy into worst recession since world war II, <https://www.worldbank.org/en/news/press-release/2020/06/08/covid-19-to-plunge-global-economy-into-worst-recession-since-world-war-ii>. Accessed 27 Jan 2021
- Yassine M, Hajj H (2010) A framework for emotion mining from text in online social networks. In: IEEE international conference on data mining workshops. Sydney, pp 1136–1142
- Zhechev V (2018) Measuring brand competitive performance—a focus on ethical brand positioning. In: Conference proceedings of 12th international days of statistics and economics. Prague, September 6–8, pp 2034–2043

# Uncertain Super-Efficiency Data Envelopment Analysis



Pejman Peykani, Jafar Gheidar-Kheljani, Donya Rahmani,  
Mohammad Hossein Karimi Gavareshki, and Armin Jabbarzadeh

**Abstract** The main goal of the current study is to propose a new method for ranking homogeneous decision-making units in the presence of uncertain inputs and/or outputs. To reach this goal, data envelopment analysis approach, super-efficiency technique, and uncertainty theory are applied. Accordingly, in this study, a novel uncertain super-efficiency data envelopment analysis approach is presented that is capable to be used under data uncertainty. Notably, the super-efficiency data envelopment analysis approach is proposed under constant returns to scale assumption and multiplier form. Additionally, to show the efficacy and applicability of the proposed method, a numerical example related to five decision-making units with two uncertain inputs and two uncertain outputs is utilized. The results indicate that the proposed uncertain super-efficiency data envelopment analysis approach is an effective and applicable method for performance evaluation and ranking of decision-making units under uncertainty environment.

**Keywords** Ranking approach · Super-efficiency · Data envelopment analysis · Uncertain data · Uncertainty theory

---

P. Peykani

School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

J. Gheidar-Kheljani (✉) · M. H. Karimi Gavareshki

Management and Industrial Engineering Department, Malek Ashtar University of Technology, Tehran, Iran

e-mail: [kheljani@mut.ac.ir](mailto:kheljani@mut.ac.ir)

M. H. Karimi Gavareshki

e-mail: [mh\\_karimi@mut.ac.ir](mailto:mh_karimi@mut.ac.ir)

D. Rahmani

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran

e-mail: [drahmani@kntu.ac.ir](mailto:drahmani@kntu.ac.ir)

A. Jabbarzadeh

Department of Automated Production Engineering, École de Technologie Supérieure (ETS), Montreal, Canada

e-mail: [armin.jabbarzadeh@etsmtl.ca](mailto:armin.jabbarzadeh@etsmtl.ca)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

311

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_19](https://doi.org/10.1007/978-3-030-85254-2_19)



## 1 Introduction

Performance assessment is one of the most important and essential real-world decision-making problems. Among the wide spectrum of performance assessment approaches, data envelopment analysis (DEA) is one of the most popular and applicable methods for performance appraisal, benchmarking, and ranking of a set of homogeneous decision-making units (DMUs) in the presence of multiple inputs and outputs (Emrouznejad et al. 2008; Liu et al. 2013,2016; Emrouznejad and Yang 2018). Despite all advantages of DEA method, it cannot be applied under data uncertainty.

Therefore, it is necessary to present a new uncertain data envelopment analysis (UDEA) approach that is capable to be employed in the presence of uncertainty (Peykani et al. 2020a). It should be noted that for proposing UDEA models, according to the types of uncertainty in data, uncertain programming approaches such as stochastic optimization (Land et al. 1993; Sueyoshi 2000; Cooper et al. 2002; Wu et al. 2013; Zha et al. 2016), fuzzy optimization (Peykani et al. 2018a,2019a,2019b,2021; Seyed Esmaeili et al. 2019; Peykani and Gheidar-Kheljani 2020), interval programming (Jahanshahloo et al. 2004; Seyed Esmaeili 2014; Rostamy-Malkhalifeh and Seyed Esmaeili 2016; Peykani and Mohammadi 2018a; Peykani et al. 2018b; Seyed Esmaeili and Rostamy-Malkhalifeh 2018), and robust optimization (Peykani and Roghanian 2015; Peykani et al. 2016,2018c,2019c,2019d,2020b; Peykani and Mohammadi 2018b) can be applied. Accordingly, variants of uncertain DEA models including stochastic DEA, fuzzy DEA, interval DEA, and robust DEA are presented by many researchers.

Accordingly, in this study, a novel uncertain super-efficiency data envelopment analysis (USEDEA) approach will be proposed to the ranking of DMUs in the presence of uncertain data. Notably, the uncertainty theory is applied for dealing with uncertain data. Moreover, to demonstrate the applicability of the proposed approach, the USEDEA is employed for performance measurement and ranking of a numerical example including five DMUs under data uncertainty.

The rest of this paper is organized as follows. The research background of the paper including super-efficiency approach as well as uncertainty theory will be introduced in Sect. 2. The modeling of uncertain super-efficiency data envelopment analysis will be presented in Sect. 3. Then, the proposed USEDEA approach is implemented for a numerical example in Sect. 4. Finally, conclusions and some directions for future research are given in Sect. 5.

## 2 Research Background

In this section, measuring super-efficiency in data envelopment analysis under constant returns to scale (CRS) and variable returns to scale (VRS) assumptions

will be discussed. Moreover, the concept and formulation of uncertainty theory for dealing with uncertain data will be explained.

### 2.1 Super-Efficiency Method

Super-efficiency DEA (SEDEA) approach was proposed by Andersen and Petersen (1993) for the first time and it is one of the most popular ranking methods in DEA literature. Please note that in SEDEA approach, the efficient DMU under evaluation can achieve a score greater than one. Therefore, all efficient DMUs can be ranked by using super-efficiency method.

Suppose that there are  $K$  homogenous  $DMU_k$  ( $k = 1, \dots, K$ ) each consuming  $I$  inputs  $x_{ik}$  ( $i = 1, \dots, I$ ) and producing  $J$  outputs  $y_{jk}$  ( $j = 1, \dots, J$ ). Also, the subscript  $p$  refers to the DMU under consideration. Moreover, the non-negative weights  $v_i$  ( $i = 1, \dots, I$ ) and  $u_j$  ( $j = 1, \dots, J$ ) are assigned to inputs and outputs, respectively. Notably, the free sign variable  $g$  allows the change of scale, and non-negative variables  $\lambda_k$  ( $k = 1, \dots, K$ ) are employed in production possibility set (PPS) formulation. Accordingly, the modeling of envelopment and multiplier forms of super-efficiency DEA approach based on CCR model (Charnes et al. 1978) (CRS assumption) are presented as Models (1) and (2), respectively:

$$\begin{aligned}
 & \text{Min} \quad \theta \\
 & \text{S.t.} \quad \sum_{\substack{k=1 \\ k \neq p}}^k x_{ik} \lambda_k - x_{ip} \theta \leq 0, \quad \forall i \\
 & \quad \sum_{\substack{k=1 \\ k \neq p}}^k \lambda_k y_{rk} - y_{rp} \geq 0, \quad \forall r \\
 & \quad \lambda_k \geq 0, \quad \forall k
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 & \text{Max} \quad \sum_{j=1}^J y_{jp} u_j \\
 & \text{S.t.} \quad \sum_{i=1}^I x_{ip} v_i = 1 \\
 & \quad \sum_{j=1}^J y_{jk} u_j - \sum_{i=1}^I x_{ik} v_i \leq 0, \quad \forall k \neq p \\
 & \quad v_i, u_r \geq 0, \quad \forall i, r
 \end{aligned} \tag{2}$$

Also, the envelopment and multiplier forms of SEDEA approach based on BCC model (Banker et al. 1984) (VRS assumption) are introduced as Models (3) and (4), respectively:

$$\begin{aligned}
 & \text{Min} \quad \theta \\
 & \text{S.t.} \quad \sum_{\substack{k=1 \\ k \neq p}}^K x_{ik} \lambda_k - x_{ip} \theta \leq 0, \quad \forall i \\
 & \quad \sum_{k=1}^K \lambda_k y_{rk} - y_{rp} \geq 0, \quad \forall r \\
 & \quad \sum_{k=1}^K \lambda_k = 1 \\
 & \quad \lambda \geq 0, \quad \forall k
 \end{aligned} \tag{3}$$

$$\begin{aligned}
 & \text{Max} \quad \sum_{j=1}^J y_{jp} u_j + g \\
 & \text{S.t.} \quad \sum_{i=1}^I x_{ip} v_i = 1 \\
 & \quad \sum_{j=1}^J y_{jk} u_j - \sum_{i=1}^I x_{ik} v_i + g \leq 0, \quad \forall k \neq p \\
 & \quad v_i, u_r \geq 0, \quad \forall i, r
 \end{aligned} \tag{4}$$

It should be noted that all the above super-efficiency DEA models are presented based on input-oriented (I-O) viewpoint and in a similar manner, output-oriented (O-O) SEDEA models can be proposed.

### 2.2 Uncertainty Theory

Let  $\Psi$  be a nonempty set and  $\Delta$  be a collection of its subsets, as a  $\sigma$ -algebra over  $\Psi$ . The pair  $(\psi, \Delta)$  is referred to as a measurable space and any member  $\varpi$  in this measurable space is called an event. An uncertain measure is a function  $\Omega : \Delta \rightarrow [0, 1]$  that satisfying the four axioms including normality, duality, subadditivity, and product. Then, the triplet  $\psi, \Delta, \Omega$  is called an uncertainty space (Wen et al. 2014; Liu and Liu 2009). For an uncertain variable  $\xi$ , the uncertainty distribution  $\Phi$  is defined as follows:

$$\Phi(z) = \Omega\{\xi \leq z\}, z \in \Re \tag{5}$$

Accordingly, linear uncertainty distribution  $\ell(a, b)$  is defined as Eq. (6) in which  $a$  and  $b$  are real numbers with  $a < b$ :

$$\Phi(z) = \begin{cases} 0, & \text{if } z \leq a; \\ \frac{z - a}{b - a} & \text{if } a \leq z \leq b; \\ 1, & \text{if } z \geq b \end{cases} \quad (6)$$

Also, the inverse uncertainty distribution of  $\ell(a, b)$  is as follows:

$$\Phi^{-1}(\alpha) = (1 - \alpha)a + \alpha b \quad (7)$$

It should be noted that uncertainty theory and uncertain measure can be applied for presenting the deterministic counterpart of uncertain chance constraints in mathematical optimization problems.

### 3 Proposed Uncertain Super-Efficiency DEA Approach

In this section, an uncertain super-efficiency DEA model will be proposed under constant returns to scale assumption and multiplier form. Notably, the inputs and outputs have a linear distribution  $\ell(\underline{x}_{ik}, \bar{x}_{ik})$  and  $\ell(\underline{y}_{jk}, \bar{y}_{jk})$ , respectively. In the first step, to consider the uncertainty on inputs and outputs, Model (2) will be converted to Model (8) as follows:

$$\begin{aligned} & \text{Max} \quad Q \\ & \text{S.t.} \quad \sum_{j=1}^J y_{jp} u_j \geq Q \\ & \quad \sum_{i=1}^I x_{ip} v_i \leq 1 \\ & \quad \sum_{j=1}^J y_{jk} u_j - \sum_{i=1}^I x_{ik} v_i \leq 0, \quad \forall k \neq p \\ & \quad v_i, u_r \geq 0, \quad \forall i, r \end{aligned} \quad (8)$$

It should be noted that the optimal solution of Model (8) is equal to Model (2) (Peykani et al. 2018d, 2019e). Now, the uncertainty measure is employed to deal with the uncertainty of data in uncertain chance constraint (UCC) as follows:

$$\begin{aligned}
 & \text{Max } Q \\
 & \text{S.t. } \Omega \left\{ \sum_{j=1}^J \tilde{y}_{jp} u_j \geq Q \right\} \geq \alpha \\
 & \Omega \left\{ \sum_{i=1}^I \tilde{x}_{ip} v_i \leq 1 \right\} \geq \alpha \\
 & \Omega \left\{ \sum_{j=1}^J \tilde{y}_{jk} u_j - \sum_{i=1}^I \tilde{x}_{ik} v_i \leq 0 \right\} \geq \alpha \quad \forall k \neq p \\
 & v_i, u_r \geq 0, \quad \forall i, r
 \end{aligned} \tag{9}$$

Then, according to Eq. (7), the deterministic counterpart of UCC in SEDEA model is presented for desired confidence level  $\alpha$  as Model (10):

$$\begin{aligned}
 & \text{Max } Q \\
 & \text{S.t. } \sum_{j=1}^J \left( (\alpha) \underline{y}_{jp} + (1 - \alpha) \bar{y}_{jp} \right) u_j \geq Q \\
 & \sum_{i=1}^I \left( (1 - \alpha) \underline{x}_{ip} + (\alpha) \bar{x}_{ip} \right) v_i \leq 1 \\
 & \sum_{j=1}^J \left( (1 - \alpha) \underline{y}_{jk} + (\alpha) \bar{y}_{jk} \right) u_j - \sum_{i=1}^I \left( (\alpha) \underline{x}_{ik} + (1 - \alpha) \bar{x}_{ik} \right) v_i \leq 0, \quad \forall k \neq p \\
 & v_i, u_r \geq 0, \quad \forall i, r
 \end{aligned} \tag{10}$$

Finally, a novel uncertain super-efficiency DEA model is proposed as Model (10) that is capable to be applied for performance assessment and ranking of peer DMUs under uncertain data with a linear distribution.

### 4 Numerical Example

In this section, the proposed USEDEA model will be evaluated by applying a numerical example. Table 1 presents the numerical example related to five DMUs with two uncertain inputs and two uncertain outputs with linear distribution:

The numerical results of the presented uncertain super-efficiency DEA model as well as ranking of DMUs at five different confidence levels including 0, 25, 50, 75, and 100%, are introduced in Tables 2 and 3, respectively:

**Table 1** Data of Five DMUs with Two Uncertain Inputs and Two Uncertain Outputs

DMUs	Inputs		Outputs	
	I (1)	I (2)	O (1)	O (2)
DMU 1	$\ell(3.5, 4.5)$	$\ell(2.5, 3)$	$\ell(2, 4)$	$\ell(4.5, 6.5)$
DMU 2	$\ell(2, 4)$	$\ell(4, 5.5)$	$\ell(1, 2)$	$\ell(3, 4)$
DMU 3	$\ell(1, 3)$	$\ell(3, 5)$	$\ell(3, 5)$	$\ell(4, 6)$
DMU 4	$\ell(4, 5)$	$\ell(4.5, 6)$	$\ell(4, 5.5)$	$\ell(2.5, 4)$
DMU 5	$\ell(3, 5)$	$\ell(2, 3.5)$	$\ell(5, 6)$	$\ell(3, 5)$

**Table 2** The Results of USEDEA Approach

DMUs	Confidence Levels				
	0%	25%	50%	75%	100%
DMU 1	3.0403	2.0436	1.3750	0.9179	0.6000
DMU 2	1.5000	0.8464	0.5444	0.3617	0.2455
DMU 3	7.2727	3.6250	2.0274	1.2020	0.7143
DMU 4	1.3750	0.9554	0.6584	0.4476	0.2979
DMU 5	4.5000	2.7857	1.8333	1.2600	0.8571

**Table 3** The Ranking of DMUs Based on USEDEA Approach

DMUs	Confidence Levels				
	0%	25%	50%	75%	100%
DMU 1	3	3	3	3	3
DMU 2	4	5	5	5	5
DMU 3	1	1	1	2	2
DMU 4	5	4	4	4	4
DMU 5	2	2	2	1	1

As it can be seen in Table 2, by increasing the confidence level  $\alpha$  from 0 to 100%, the results of USEDEA model are decreased. Also, the effect of data uncertainty on the ranking of DMUs can be clearly observed in Table 3.

## 5 Conclusions

In this study, a new approach is introduced for ranking of DMUs under uncertain data. For proposing this approach, DEA model, super-efficiency method, and uncertain measure are utilized. Also, by applying a numerical example, the implementation of the presented super-efficiency DEA model is illustrated. Notably, the results show the

efficacy of the proposed USEDEA model for fully ranking of DMUs in the presence of uncertain data. For future studies, the uncertain super-efficiency DEA model can be combined with machine learning models as an expert system for the prediction of data and consequently, future performance measurement of DMUs (Hong et al. 1999; Yeh et al. 2010; Li et al. 2017; Shahhosseini et al. 2019a,2019b; Salehi et al. 2020). Additionally, the USEDEA approach can be developed for other uncertain distribution such as zigzag and normal.

## References

- Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. *Manage Sci* 39(10):1261–1264
- Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manage Sci* 30(9):1078–1092
- Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. *Eur J Oper Res* 2(6):429–444
- Cooper WW, Deng H, Huang Z, Li SX (2002) Chance constrained programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis. *J Operat Res Soc* 53(12):1347–1356
- Emrouznejad A, Yang GL (2018) A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socioecon Plann Sci* 61:4–8
- Emrouznejad A, Parker BR, Tavares G (2008) Evaluation of research in efficiency and productivity: a survey and analysis of the first 30 years of scholarly literature in DEA. *Socioecon Plann Sci* 42(3):151–157
- Hong HK, Ha SH, Shin CK, Park SC, Kim SH (1999) Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning. *Expert Syst Appl* 16(3):283–296
- Jahanshahloo GR, Hosseinzadeh Lotfi F, Moradi M (2004) Sensitivity and stability analysis in DEA with interval data. *Appl Math Comput* 156(2):463–477
- Land KC, Lovell CK, Thore S (1993) Chance-constrained data envelopment analysis. *Manag Decis Econ* 14(6):541–554
- Li Z, Crook J, Andreeva G (2017) Dynamic prediction of financial distress using Malmquist DEA. *Expert Syst Appl* 80:94–106
- Liu B, Liu B (2009) *Theory and practice of uncertain programming*. Springer, Berlin
- Liu JS, Lu LY, Lu WM, Lin BJ (2013) A survey of DEA applications. *Omega* 41(5):893–902
- Liu JS, Lu LY, Lu WM (2016) Research fronts in data envelopment analysis. *Omega* 58:33–45
- Peykani P, Gheidari-Kheljani J (2020) Performance appraisal of research and development projects value-chain for complex products and systems: the fuzzy three-stage DEA approach. *J New Res Math* 6(25):41–58
- Peykani P, Roghanian E (2015) The application of data envelopment analysis and robust optimization in portfolio selection problem. *J Operat Res Appl* 12(44):61–78
- Peykani P, Mohammadi E, Jabbarzadeh A, Jandaghian A (2016) Utilizing robust data envelopment analysis model for measuring efficiency of stock, a case study: Tehran stock exchange. *J New Res Math* 1(4):15–24
- Peykani P, Seyed Esmaeili FS, Rostamy-Malkhalifeh M, Hosseinzadeh Lotfi F (2018a) Measuring productivity changes of hospitals in Tehran: the fuzzy Malmquist productivity index. *Int J Hospital Res* 7(3):1–17

- Peykani P, Mohammadi E, Pishvae MS, Rostamy-Malkhalifeh M, Jabbarzadeh A (2018d) A novel fuzzy data envelopment analysis based on robust possibilistic programming: possibility, necessity and credibility-based approaches. *RAIRO-Operat Res* 52(4–5):1445–1463
- Peykani P, Mohammadi E, Hosseinzadeh Lotfi F, Tehrani R, Rostamy-Malkhalifeh M (2019a) Performance evaluation of stocks in different time periods under uncertainty: fuzzy window data envelopment analysis approach. *Financ Eng Secur Manag* 10(40):304–324
- Peykani P, Mohammadi E, Rostamy-Malkhalifeh M, Hosseinzadeh Lotfi F (2019b) Fuzzy data envelopment analysis approach for ranking of stocks with an application to Tehran stock exchange. *Adv Math Financ Appl* 4(1):31–43
- Peykani P, Mohammadi E, Seyed Esmaeili FS (2019d) Stock evaluation under mixed uncertainties using robust DEA model. *J Qual Eng Prod Optim* 4(1):73–84
- Peykani P, Mohammadi E, Emrouznejad A, Pishvae MS, Rostamy-Malkhalifeh M (2019e) Fuzzy data envelopment analysis: an adjustable approach. *Expert Syst Appl* 136:439–452
- Peykani P, Mohammadi E (2018) Interval network data envelopment analysis model for classification of investment companies in the presence of uncertain data. *J Indus Syst Eng* 11(Special issue: 14th international industrial engineering conference):63–72
- Peykani P, Mohammadi E (2018) Robust data envelopment analysis with hybrid uncertainty approaches and its applications in stock performance measurement. In: *The 14th international conference on industrial engineering, Iran*
- Peykani P, Mohammadi E, Seyed Esmaeili FS (2018) The classification of investment companies using the interval network data envelopment analysis model. In: *The 14th international conference on industrial engineering, Iran*
- Peykani P, Mohammadi E, Sadjadi SJ, Rostamy-Malkhalifeh M (2018) A robust variant of radial measure for performance assessment of stock. In: *The 3th international conference on intelligent decision science, Iran*
- Peykani P, Seyed Esmaeili FS, Hosseinzadeh Lotfi F, Rostamy-Malkhalifeh M (2019) Estimating most productive scale size in DEA under uncertainty. In: *The 11th national conference on data envelopment analysis, Iran*
- Peykani P, Mohammadi E, Farzipoor Saen R, Sadjadi SJ, Rostamy-Malkhalifeh M (2020) Data envelopment analysis and robust optimization: a review. *Expert Syst* 37(4):e12534
- Peykani P, Mohammadi E, Jabbarzadeh A, Rostamy-Malkhalifeh M, Pishvae MS (2020) A novel two-phase robust portfolio selection and optimization approach under uncertainty: a case study of Tehran stock exchange. *PLoS One* 15(10):e0239810
- Peykani P, Mohammadi E, Emrouznejad A (2021) An adjustable fuzzy chance-constrained network DEA approach with application to ranking investment firms. *Expert Syst Appl* 166:113938
- Rostamy-Malkhalifeh M, Seyed Esmaeili FS (2016) Computing the efficiency interval of decision making units (DMUs) having interval inputs and outputs with the presence of negative data. *J New Res Math* 1(4):5–14
- Salehi V, Veitch B, Musharraf M (2020) Measuring and improving adaptive capacity in resilient systems by means of an integrated DEA-Machine learning approach. *Appl Ergonom* 82:102975
- Seyed Esmaeili FS (2014) The efficiency of MSBM model with imprecise data (interval). *Int J Data Envelop Anal* 2(1):343–350
- Seyed Esmaeili FS, Rostamy-Malkhalifeh M (2018) Using interval data envelopment analysis (IDEA) to performance assessment of hotel in the presence of imprecise data. In: *The 3th international conference on intelligent decision science, Iran*
- Seyed Esmaeili FS, Rostamy-Malkhalifeh M, Hosseinzadeh Lotfi F (2019) The possibilistic Malmquist productivity index with fuzzy data. In: *The 11th national conference on data envelopment analysis, Iran*
- Shahhosseini M, Hu G, Archontoulis SV (2020) Forecasting corn yield with machine learning ensembles. *Front Plant Sci* 11:1120
- Shahhosseini M, Hu G, Pham H (2019) Optimizing ensemble weights for machine learning models: a case study for housing price prediction. In: *INFORMS international conference on service science*. pp. 87–97. Springer, Cham



- Shahhosseini M, Martinez-Feria RA, Hu G, Archontoulis SV (2019) Maize yield and nitrate loss prediction with machine learning algorithms. *Environ Res Lett* 14(12):124026
- Sueyoshi T (2000) Stochastic DEA for restructure strategy: an application to a Japanese petroleum company. *Omega* 28(4):385–398
- Wen M, Guo L, Kang R, Yang Y (2014) Data envelopment analysis with uncertain inputs and outputs. *J Appl Math* 307108
- Wu C, Li Y, Liu Q, Wang K (2013) A stochastic DEA model considering undesirable outputs with weak disposability. *Math Comput Model* 58(5–6):980–989
- Yeh CC, Chi DJ, Hsu MF (2010) A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Syst Appl* 37(2):1535–1541
- Zha Y, Zhao L, Bian Y (2016) Measuring regional efficiency of energy and carbon dioxide emissions in China: a chance constrained DEA approach. *Comput Oper Res* 66:351–361

# Using Data Mining Techniques for Designing Patient-Friendly Hospitals



İpek Deveci Kocakoç and Gökçe Baysal Türkölmez

**Abstract** Spending a long time in the hospital and the intense circulation between the clinics involve various risks for both patients and medical officials in the hospital. Especially in the COVID-19 outbreak, the minimization of this period is vital because of reducing the risk of transmission. Symptomatic patients can be immediately taken into the isolation and treatment process, while asymptomatic patients continue to spread the disease. Because patients suffering from other diseases like diabetes, cancer, or other chronic illness also have immune system problems, epidemics can be fatal for them. Therefore, especially during the pandemic, patients tend to delay their hospital visits due to the risk of contamination. In this situation, unless they reach the diagnosis and treatment, they will face more serious health problems. If we provide patients uncrowded hospitals and shorter hospital visits, we will reduce the risk of transmission of COVID-19 and other seasonal epidemics. Therefore, it is necessary to determine which clinics are visited frequently by the patients and which clinics and medical units work together in the diagnosis and treatment process. In this chapter, data mining techniques used in healthcare system design are explained and exemplified by a real-life case study. Analyzing patient data by using data mining techniques allows us to reach the aim of this chapter. Association rules between clinics and other related medical units like blood-letting and nuclear medicine services are determined. They also reveal the circulation of patients in the hospital. Frequency analysis shows crowded clinics and other medical units. Minimizing this circulation and crowd of patients in the hospital also minimizes the risk of transmission of COVID-19. In this chapter, six months' data of patients treated in a hospital in Turkey are used. These data include demographic information of the patients, as well as which clinics they visited and how many days they were treated in the hospital. As a result of the data mining analysis, clinics and medical units working together in the diagnosis and treatment process and the most crowded clinics will be determined. Recommendations will be made to reduce the distance between the clinics and units

---

İ. D. Kocakoç (✉) · G. B. Türkölmez  
Econometrics Department, Dokuz Eylül University, Izmir, Turkey  
e-mail: [ipek.deveci@deu.edu.tr](mailto:ipek.deveci@deu.edu.tr)

G. B. Türkölmez  
e-mail: [gokce.baysal@deu.edu.tr](mailto:gokce.baysal@deu.edu.tr)

which have associations and increase the service capacities of the most crowded clinics, respectively. Thus, the application of data mining techniques for designing patient-friendly healthcare services is presented.

## 1 Introduction

Pandemics made important changes to the lifestyle in history. For avoiding transmission of disease which causes pandemics, new rules need to be applied. Because effects of a pandemic will be seen for years, those rules cause to design of new systems and processes for buildings and structures which people have to work together. They may also prevent new diseases to transform pandemics.

Physical distancing has crucial importance during the pandemic of Covid-19 which is a highly contagious viral infection. The virus was commenced from the Wuhan seafood market of China at the end of 2019. It has been spreading around the world rapidly due to onward transmission. It has caused respiratory infections and Severe Acute Respiratory Syndrome (SARS) in humans (Mohapatra et al. 2020). In recent data, more than 2 million people died due to Covid-19 and approximately 100 million people have been infected according to the World Health Organization. Viruses have the ability of mutation for adapting to new conditions. Covid-19 has some mutations called the variant. Because a new variant of Covid-19 is more contagious, designing a new system for increasing the distance amongst people becomes more urgent and important.

Hospitals are crowded buildings where people work and visit 24 h a day. They need to be evaluated and investigated in terms of medical officials and patients. Medical officials have to be exposed to viruses and other kinds of infectious diseases because of patients. They need to use personal protective equipment (PPE) and work in a wide and isolated area. Patients visit a hospital because of diseases that need to be treated. The disease may be contagious and it may cause a pandemic. For example, during Covid-19 there are two types of patients; symptomatic and asymptomatic. The symptomatic patients can be isolated and included a treatment process immediately, but the asymptomatic patients spread Covid-19 without any sign. On the other hand, patients suffering from diseases like diabetes, cancer, cardiovascular problems, or other chronic illness may have some immune system problems. If they are infected by Covid-19, their diseases may be worse or they may die. Therefore, they need to be isolated from other patients, especially asymptomatic Covid-19 patients. This kind of isolation can be provided by shorter hospital visits in terms of time and distance.

Hospital visits in a shorter time are related to the shorter distance between clinics in the hospital and shorter waiting times for appointments. Shorter waiting times which are not a subject of this study are about scheduling the appointment and the number of doctors working in the related polyclinic. A shorter distance between clinics decreases the traveling time in the hospital and the possibility of being exposed to infections. In this study, we have focused on this problem and recommended a system design to provide a patient-friendly hospital.

For designing a patient-friendly hospital, firstly we need to analyze patient data. In this context, we have six months' data, taken from a hospital in Turkey, including demographics of patients, which clinics are visited, and duration of treatments. The hospital provides comprehensive treatment services with 48 clinics and it is placed in the area of 46000m<sup>2</sup>. In this study, data analytics consists of popular data mining algorithms for association rules and frequent itemset mining. Data mining is a process of knowledge extraction by pattern discovery in a huge amount of data called big data (Jothi and Husain, 2015). The data mining application aims to reveal Association Rules between clinics and to find clinics that are visited frequently. New pathways between clinics that are visited together in a treatment process and some developments and improvements for frequently visited clinics are recommended as results of the data analytics.

The parts of this chapter are organized as follows: the literature review is presented in Sect. 2. The concept of data mining and the application area of data mining algorithms are clarified in a general perspective in Sect. 3. The association rule mining and frequent itemset mining used in this study are explained in a detailed manner in Sects. 3.1 and 3.2, respectively. Explanation about algorithms is presented in Sect. 3.3. Three main algorithms of interest, Apriori, FP-growth, and Eclat, are also given. Then the analysis and results are given in Sect. 4. Conclusion and future research are placed as the last part of the chapter.

## 2 A Brief Literature Review of Data Mining on Health care

Articles published in academic journals, books, and dissertations are evaluated in almost 20 years period in this study. Because technological improvements and new approaches to data analytics have changed the concept of data through big data, we focus on this period between 2000 and 2021 to provide a perspective for application areas and a theoretical framework. This review has some important inferences for healthcare systems and healthcare management for the future studies.

Data mining techniques are consist of supervised and unsupervised techniques in machine learning like association, classification, clustering, pattern discovering, and data visualization. In recent years, meta-rule-guided mining is also included in those techniques (Jothi and Husain 2015). There are various techniques and methods used in data analytics whereas their usage in the area of health care is generally separated into four different domains. One of them is analyzing data of a particular disease or treatment and providing some inferences like predictions about the frequency of this disease in the future and effects of treatment on this disease by using classification algorithms. The other one is the analysis of public health data for infections/pandemics control surveillance. Papers about the frequency of diseases in a particular city, the period of time, or group of people are contained in this group. The next one focuses on health care and medical resources management, healthcare information systems, and hospital management (Sharma and Mansotra 2014). The

last one presents a framework about the usage of data mining methods to health care and medical data.

In the first area, a lot of papers about a particular disease or treatment are analyzed by data mining. Especially, diabetes, heart diseases, kidney illness, and cancer have been focused on as application context in the literature. Kaur and Wasan (2006) examined classification-based data mining methods like decision trees and artificial neural networks on diabetes data. Contributions of sick and health factors to heart disease in females and males were revealed by using association rule mining (Nahar et al. 2013). Huang (2013) investigated the correlation between disease and abnormal test results by using association rules. Findings are used to set a disease-prevention knowledge database for assisting healthcare providers in diagnosing process. A system was developed by using association rule mining methods to find a correlation between heart diseases as primary and secondary (Rashid et al. 2014). Care targets for hospitalized dementia patients were analyzed by association rules to determine the needs of patients clearly. Therefore, the management of a hospital could provide better health care to them (Shih et al. 2018).

Predicting diseases by using data mining techniques on medical data was another subject of this area. Classification and ensemble classifiers were used to predict hypertension as a decision attribute with eight diseases including diabetes, acute myocardial infarction, pneumonia, cardiac dysrhythmias, and other five diseases as condition attributes (Huang et al. 2012). Locally frequent diseases were investigated and presented by the Apriori algorithm and visualization techniques, respectively. A prototype application was developed to show the efficiency of the method (Khaleel et al. 2013). Data mining methods were applied to develop a prediction system for decision support in a hospital in Lebanon. Data from the healthcare information system was processed by dynamic rough sets attribute reduction with multi classifier random forest. The purpose of this analysis was to increase service quality and to decrease the cost of treatment in the hospital. The data mining approach was applied to four different medical cases like acute appendicitis, premature birth, osteoporosis, and coronary heart disease. Web-based interfaces were developed to determine the risk level of medical processes for patients (Shahin et al. 2014). The frequency of diseases was identified by an association rule-based Apriori algorithm in a particular region for a particular period of time (Ilayaraja and Meyyappan 2013; Hande et al. 2015).

Data mining analyses about pharmacies, medicines, and other medical equipment and systems as medical resources provide important inferences to develop better healthcare systems for patients. Relationships between procedures like medical tests on patients and various diagnoses reported were found by using association rules (Doddi et al. 2001). Data mining and artificial neural network were used to predict the length of stay in the hospital of veteran patients who suffer from spinal cord injury. The results of this analysis were used to decrease cost by effectiveness in resource allocation (Kraft et al. 2003). Nursing diagnosis data were used to build a recommender system for developing nursing care plans by itemset mining (Duan et al. 2011). Electronic healthcare database provides data about side effects of drugs.

A concept method for the refining side effect of drugs was developed by using association rule mining (Reps et al. 2014). A robotic prescription dispensing system and a planogram were developed for the pharmacy automation system. The associations in the prescribed medication regime were revealed by using FP-growth as one of the Association rule mining algorithms (Khader et al. 2016). The hospital information system has different medical subsystems. If there is a problem with data sharing or data storage, this may cause that these data couldn't be processed properly. An association rule mining method, including multiple minimum supports, was proposed to find association rules from the radioimmunoassay data by using laboratory information system and departmental registration system of a hospital. This analysis produced important results about improving doctor-patient relationships and healthcare quality (Lin et al. 2016). Surveillance of drugs and relationships between drugs were analyzed by the network analysis and the association rule mining. A target drug network was defined and drugs taken together were determined in this study (Belyi et al. 2016). For recommending suitable medicine to a particular disease, the relationship between the medicine and the disease was investigated by k-mean clustering to classify diseases and the Apriori algorithm as the association rule mining (Harahap et al. 2018).

Some studies presented frameworks about the usage of methods and design process of data mining analysis on medical data (Kaur and Wasan 2006; Hande et al. 2015; Țăranu 2016). Swathi and Prajna (2016) researched the utility of different data mining methods like bunching, relapse, affiliation, and arrangement on medical data. Birnbaum (2004) identified the data mining techniques used in the healthcare data. Canlas (2009) gave examples of current issues and applications of data mining in health care.

### 3 Association Rule Mining

Data mining reveals meaningful information from the complexity of data (Ogundele et al. 2018). It consists of data analysis and knowledge discovery. In this manner, data mining includes techniques like visualization, database design, artificial intelligence, machine learning, and pattern recognition to analyze big data (Yoo et al. 2012). Companies and industries often use the data mining as a decision-making tool to identify relationships amongst items predicting customer buying behavior. Hospitals apply data mining to patient data for improving service quality, predicting diseases, and providing disease and treatment surveillance. In the healthcare application, relationships between patients and disease, patients and medicine, disease and treatment are often investigated to discover intrinsic knowledge on the dataset.

Methods in data mining are evaluated in three main subjects: classification, association rule, and clustering (Mahindrakar and Hanumanthappa 2013). In this study, we aimed to define relationships between clinics and to find the clinics frequently visited. The analyses provide us to plan a redesign of the hospital layout considering clinics visited together and to make some developments for crowded clinics.

Association Rule Mining discovers unique rules and unexpected relationships from big datasets. It is usually used to analyze the market-basket problems. The market database contains sales data with transaction records for each customer and it is a large collection of itemset about sales and customers. Association rules reveal information about customer purchasing behavior which includes the items bought together and discover the most frequent item in the itemset (Khader et al. 2016). Basic concepts of association rule mining can be explained as follows:

Items are introduced as goods like milk, bread, or any kind of product purchased in a store.  $I = \{i_1, i_2, \dots, i_m\}$  is a set of items.  $X$  is called itemset and it is a subset of  $I$ . Transaction database is identified by  $T = \{t_1, t_{i+1}, \dots, t_n\}$  where  $T \subseteq I$ . Each transaction  $t$  has a *tid* (transaction identifier) and a *t-itemset*;  $t = (tid, t - itemset)$  (Zhang and Zhang 2003). An association rule is showed as  $\{i_p, i_q\} \Rightarrow \{i_k\}$ ; if a customer purchased items  $i_p$  and  $i_q$  then he will probably buy  $i_k$  (Belyi et al. 2016). If  $\{i_p, i_q\}$  is represented by A and  $\{i_k\}$  is represented by B, association rule is a pair (A, B) of itemset. This relationship is denoted by  $A \rightarrow B$  and A is the antecedent and B is the consequent of the rule  $A \rightarrow B$ . It means that after A, B happens (Ramezankhani et al. 2015). Two important metrics called *support* and *confidence* are used in the association rule mining. The support is the ratio of transactions in T containing X.

$$supp(X) = |X(t)|/|T|$$

where  $X(t) = \{tinT|tcontainsX\}$ . The support of a rule  $X \rightarrow Y$  is demonstrated as  $X \cup Y$ , where X and Y are itemsets, and it is the relative frequency of transactions containing X and Y (Zhang and Zhang, 2003). The number of occurrences of an itemset in the databases also determines the support of it. The number of occurrences is demonstrated by *the count* value in the analysis (Hipp et al. 2000). The confidence of a rule is the conditional probability which is formulated below (Belyi et al. 2016):

$$conf(X \rightarrow Y) = \frac{|(X \cup Y)(t)|}{|X(t)|} = \frac{supp(X \cup Y)}{supp(X)} = P(Y|X)$$

Confidence ( $conf(X \rightarrow Y) = P(Y|X)$ ) is the probability of finding consequent (Y) in transactions under the condition of containing antecedent (X) of the same transactions (Hornik et al. 2005). The decision-maker determines thresholds for support and confidence, called *minsupp* and *minconf*, respectively. If values of the support and the confidence are greater and equal to the thresholds, then this rule  $X \rightarrow Y$  is identified as valid.

If a huge number of association rules, which are higher than *minsupp* and *minconf*, are generated, then additional interest measures like *lift* are required to use to rank or filter the rules.

$$lift(X \Rightarrow Y) = supp(X \cup Y)/(supp(X)supp(Y))$$

The lift is the deviation of all rules from expected support according to the supports of  $X$  as antecedent and  $Y$  as consequent which are given independently. Stronger associations are demonstrated by higher lift values (Hornik et al. 2005).

Other important measures are coverage and length. Major patterns in the dataset are explored by coverage value. In that manner, maximum or higher values of coverage point out the major patterns. Because short rules are easily understood, minimum length values are preferable (Zielosko 2016).

### 3.1 Frequent Itemset Mining

Frequent itemset mining is one of the most popular data mining methods which is developed for market-basket analysis such as association rule mining (Borgelt 2012). Frequent itemset mining is a special case of association rule mining. If values of *minsupp* and *minconf* are given, each frequent itemset  $X$  represents rule  $X \rightarrow \{\}$  with 100% confidence. In the same way, the support of the rule is the support of  $X$ . For every frequent itemset  $I$ , all rules  $X \rightarrow Y$ , with  $X \cup Y = I$ , hold with *minconf*. If the support of the rule is bigger than *minsupp*, the association rule is *frequent* (Goethals 2003).

### 3.2 Algorithms

Three main algorithms, Apriori, FP-growth, and Eclat, are used to find the association rules and frequent itemsets in the area of data mining. They are explained in the continuation of this part. In the process of determining association rules and finding frequently visited clinics for designing a patient-friendly hospital, Apriori and Eclat algorithms are used for frequent itemsets, Apriori also generates association rules (Hornik et al. 2005).

#### 3.2.1 Apriori Algorithm

Apriori is the most important and influential algorithm of association rule mining. It finds frequent itemset of Boolean association rules (Liu 2010). Apriori is an iterative algorithm which is used breadth-first search and counts transactions (Hornik et al. 2005). The database is scanned by multiple passes in this algorithm. In Breadth-first search  $k$ -itemsets are used to discover  $(k + 1)$ -itemsets (Ilayaraja and Meyyappan 2013).

Apriori algorithm uses a horizontal database layout (Saxena and Gadhiya 2014). In the solution process of the Apriori algorithm, the problem is evaluated into two subproblems. Firstly, frequent itemsets which have transaction support bigger than



or equal to the *minsupp* are found. According to the frequent itemsets, all the association rules are generated. The confidence ratio which is explained above, for each association rule is calculated. The association rule holds only the confidence ratio is bigger than or equal to the *minconf*. The pseudocode of the Apriori algorithm is given below (Agrawal et al. 1996).

```

procedure AprioriAlg()
begin
   $L_1 := \{frequent\ 1 -\ itemset\}$ ;
  for ( $k := 2; L_{k-1} \neq \emptyset; k++$ ) do {
     $C_k := apriori - gen(L_{k-1}); //New\ candidates$ 
    forall transactions  $t$  in the dataset do {
      forall candidate  $c \in C_k$  contained in  $t$  do
         $c.count++$ ;
    }
     $L_k := \{c \in C_k | c.count \geq minsupp\}$ 
  }
   $Answer := \bigcup_k L_k$ ;
end

```

### 3.2.2 FP-Growth Algorithm

The FP-growth (Frequent Pattern Growth) algorithm is the fastest algorithm to find frequent itemsets because it uses a lesser number of scans of the database than the Apriori algorithm. It reaches the frequent itemsets only in two scans of the database (Zhang et al. 2008). It uses a projected database layout (Saxena and Gadhiya 2014). The process of the FP-growth algorithm is introduced as follow (Borgelt 2005):

- The frequencies of items are calculated by an initial scanning. A value of *minsupp* is determined by the decision-maker for comparing items. Infrequent items which are smaller than *minsupp* are removed from the transaction database.
- The items of each transaction are sorted in descending order according to their frequencies.
- The reduced database is obtained.

As a detailed explanation, the pseudocode of the FP-growth algorithm is given below (Goethals 2003).

```

Input:  $\mathcal{D}, \text{minsupp}, I \subseteq \mathcal{I}$ 
Output:  $\mathcal{F}[I](\mathcal{D}, \text{minsupp})$ 
 $\mathcal{F}[I] := \{\}$ 
for all  $i \in \mathcal{I}$  occurring in  $\mathcal{D}$  do
   $\mathcal{F}[I] := \mathcal{F}[I] \cup \{I \cup \{i\}\}$ 
  // Create  $\mathcal{D}^i$ 
   $\mathcal{D}^i := \{\}$ 
   $H := \{\}$ 
  for all  $i \in \mathcal{I}$  occurring in  $\mathcal{D}$  such that  $j > i$  do
    if  $\text{support}(I \cup \{i, j\}) \geq \text{minsupp}$  then
       $H := H \cup \{j\}$ 
    end if
  end for
  for all  $(tid, X) \in \mathcal{D}$  with  $i \in X$  do
     $\mathcal{D}^i := \mathcal{D}^i \cup \{(tid, X \cap H)\}$ 
  end for
  // Depth-first recursion
  Compute  $\mathcal{F}[I \cup \{i\}](\mathcal{D}^i, \text{minsupp})$ 
   $\mathcal{F}[I] := \mathcal{F}[I] \cup \mathcal{F}[I \cup \{i\}]$ 
end for

```

### 3.2.3 Eclat Algorithm

Eclat (Equivalence Class Clustering and bottom-up Lattice Traversal) is one of the frequent itemset mining algorithms and employs depth-first search. Eclat algorithm uses a total order on the items like Apriori and vertical database layout (Schmidt-Thieme 2004). The intersection-based approach is applied for determining the support of an itemset (Goethals 2003). Depth-first recursion is provided by a prefix  $I$  which is input parameters and these parameters differentiate Eclat from Apriori. The prefix pattern, determined by the prefix  $I$ , must be existed in any itemsets found out by the Eclat algorithm. In the initial run of Eclat, a prefix doesn't need to be specified. All of the single-item frequent itemsets are explored by this initial (Heaton 2016). Pseudocode of the Eclat algorithm is given below (Goethals 2003).

```

Input:  $\mathcal{D}, \text{minsupp}, I \subseteq \mathcal{I}$ 
Output:  $\mathcal{F}[I](\mathcal{D}, \_)$ 
 $\mathcal{F}[I] := \{\}$ 
for all  $i \in \mathcal{I}$  occurring in  $\mathcal{D}$  do
   $\mathcal{F}[I] := \mathcal{F}[I] \cup \{I \cup \{i\}\}$ 
  // Create  $\mathcal{D}^i$ 
   $\mathcal{D}^i := \{\}$ 
  for all  $i \in \mathcal{I}$  occurring in  $\mathcal{D}$  such that  $j > i$  do
     $C := \text{cover}(\{i\}) \cap \text{cover}(\{j\})$ 
    if  $|C| \geq \text{minsupp}$  then
       $\mathcal{D}^i := \mathcal{D}^i \cup \{(j, C)\}$ 
    end if
  end for
end for
// Depth-first recursion
Compute  $\mathcal{F}[I \cup \{i\}](\mathcal{D}^i, \text{minsupp})$ 
 $\mathcal{F}[I] := \mathcal{F}[I] \cup \mathcal{F}[I \cup \{i\}]$ 
end for

```

### 3.3 Software and Tools

Data mining algorithms are applied to datasets by using some particular software and tools which are designed for specific problems or dataset structures. There is a lot of software and tools for association rules and frequent itemset mining, which use related algorithms explained above. They are separated into two classes as free and commercial ones which are listed in Table 1.

## 4 Data and Analysis

Association mining is often used in the medical field to explore associations between symptoms and illnesses, or between treatments and adverse reactions. In this study, it is used to find out the associations between clinics of a private hospital. In other words, our purpose is to create association rules for a database of clinics attended by patients during a single hospital stay.

The dataset consists of patient data of a private hospital during the 1.1.2018–1.9.2019 period. The original dataset shows all patient visit data, including daily visits to clinics, ER, and inpatient data in 801,726 lines of visit data for 239,574 unique patients. Since our purpose is to associate clinic visits, we filtered the data to include only daily clinic visits, which is 589800 visits for 190,442 unique patients. The frequency of clinic visits is given in Table 2.

**Table 1** Tools and software used for data mining

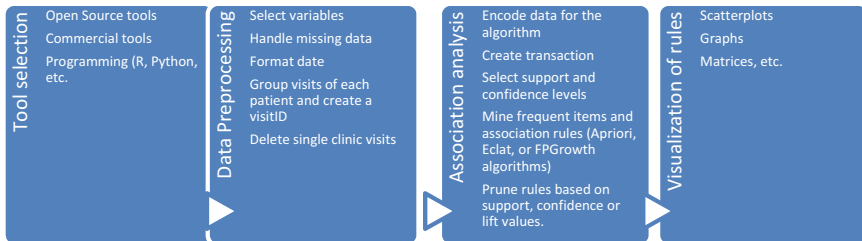
Tool	F/C	Findings and operations	Algorithms
Arules R package	Free	Association rules, frequent itemset, data representation, data manipulation, and analyzing transaction data and patterns	Apriori and Eclat
ARMiner project	Free	Association Rules	Apriori, FP-growth, Closure algorithm (ClosureOpt)
IBM SPSS modeler	Commercial	Data analytics and data mining	Decision tree, machine learning algorithms, Hadoop, Spark
KNIME	Commercial	Data analytics, workflow reports, data mining	Machine learning algorithms, Hadoop and Spark
LPA data mining toolkit	Commercial	Association rules, frequent itemsets	Artificial intelligence, knowledge-based decision support, expert systems
OPUS miner	Free	Association rules, frequent itemsets	OPUS search algorithm
Orange data mining	Free	Classification, text mining, clustering, visualization, predictive models	Machine learning algorithms
RapidMiner:	Commercial	Creating predictive models	Machine learning algorithms, TensorFlow, Hadoop, and Spark
XLMiner	Commercial	Association rules, frequent itemsets data visualization, forecasting, data mining, text mining	Regression, decision trees, neural networks, ensembles, PMML (Predictive Model Markup Language)
Wizsoft WizRule	Commercial	Association rules, frequent itemsets	If-then rules and mathematical formulations

The most frequently visited clinic is Gynecology and Obstetrics, followed by Cardiology and Children’s Health and Diseases clinics. This information leads us to position these frequently visited clinics to easily reachable locations in the hospital. However, in order to make an arrangement to shorten the travel distance according to clinics visited at one single visit, we need to know which clinics are associated. Association analysis helps us reach this goal.

The roadmap of this study is given in Fig. 1. Although some steps could change based on the application area, this roadmap is valid for most association analysis studies.

**Table 2** Frequency of daily clinic visits

Clinic	# Visits	Clinic	# Visits
Obstetrics and gynecology	67,118	Neurosurgery	8173
Cardiology	44,031	Family and community medicine	6365
Pediatrics	40,350	Pediatric allergy and immunology	5773
Dermatology	34,780	Immunology, and allergy medicine (pulmonary)	5327
Clinical oncology	32,017	Infectious diseases	4957
Otolaryngology	29,791	Cardiovascular surgery	4629
Psychiatry	27,299	Anesthesiology	4373
Ophtalmology (and visual sciences)	23,621	Nefrology	3519
Orthopaedic surgery(and traumatology)	23,265	Occupational medicine	3190
General internal medicine	22,940	Plastic surgery	2993
Urology	21,985	Radiology	2753
Biochemistry	20,076	Pediatric surgery	2676
Endocrinology, diabetes, and metabolism	19,468	Radiation oncology	2487
Surgery	17,921	Nutrition and dietetics	2366
Neurology	17,650	Thoracic surgery	1220
Physical medicine and rehabilitation	17,436	Laboratory medicine (clinical pathology)	1218
Pulmonary medicine	16,980	Psychology	484
General practice	14,766	Pediatric hematology and oncology	463
Gastroenterology	14,471	Perinatology	103
Rheumatology	10,503	Neonatology	31
Hematology	10,222		



**Fig. 1** Association analysis roadmap for the study

## 4.1 Tool Selection

In this chapter, the “arules” library in R is used with some additional coding, but the open-source or commercial tools described in Sect. 3.3 can also be utilized. The codes and a small dataset are given in <https://github.com/ipekdk/Hospital-clinic-visits-association-analysis>. Arules library is developed by Hahsler et al. (2005) to mine frequent itemsets, maximal frequent itemsets, closed frequent itemsets, and association rules, and the latest version of 1.6–6 is published on 15.05.2020 for visualization of rules, arulesViz library (Hahsler 2017; Hahsler and Chelluboina 2019), an extension of “arules” with various visualization techniques for association rules and itemsets, is used.

## 4.2 Data Preprocessing

Before making the analysis, the data needs preprocessing. The first step of preprocessing is variable selection. Association analysis needs at least the patient ID, visit date, and visited clinic columns for each row (visit). If different variations of analysis are aimed at, demographic information of patients will also be required. Our dataset also includes the gender, insurance type, and birth year of each patient. Patient data is completely non-specific and blind. Custom association analysis doesn’t require any personal patient information; therefore, it complies with personal data protection law. However, if detailed analysis based on specific and full demographic information is needed, personal data protection should be considered (Table 3).

The second step of preprocessing is handling missing data. To perform any data mining model, we should handle the NaN or Null values in our datasets. Since the data is from the hospital’s database, and all variables of interest are compulsory fields in the database, there is no missing data in this case. However, there might be some missing data in other datasets and those missing data can damage the analysis. Deleting rows (visits) with missing data is one option that has the risk of losing data points with valuable information. Imputing or fixing the data by connecting it with another table in the database are other options. Since handling missing data for data

**Table 3** Attributes of the dataset used in analyses

Attribute name	Explanation	Data type
patientID	ID number of the patient	Numeric
birthyear	Birth year of the patient	Numeric
gender	Gender of the patient	Categorical
submission_date	Date of visit	Date
Clinic	Name of the visited clinic	Categorical
insurance	Type of insurance	Categorical

mining is not in the scope of this chapter, the reader may refer to Badr (2019) and SCI2S (2021) for recent information.

Date formatting is the third step of preprocessing. The date format retrieved from the database should be suitable for association mining. If the aim is the find concurrent events, the date or timestamp of each event should be in a format to distinguish all events occurring at the same time frame. For this hospital case, the date was formatted like “2018-01-01 14:41:00” with the timestamp of the data entry to the hospital’s database. Since visits can be one after another in a day, the hour:minute: second information in date format makes it impossible to determine the concurrency of visits of a patient in a day. The date is formatted only to include year-month-day information. For other cases, formatting to have the exact hour, minute, or second of the events might be needed.

The fourth step is to group visits of each patient and create a visitID. This visitID shows the visits of an individual patient in one day. By using this variable, the clinics visited by a patient in one day can be determined. This variable stands for “transaction ID” in the classical market-basket analysis.

The last step of preprocessing is deleting visits (and patients) that are made to only one clinic in a day. Since the aim is to find the associations between clinic visits, single visits will not be of use and should be removed from the dataset. After removing single visits, 135,723 multiple visit data are available for analysis. The original and preprocessed data are given in Fig. 2a, b.

Other preprocessing steps may be needed for different datasets or mining tools. Negative values or very small values may be eliminated.

### 4.3 Frequent Itemset Mining and Association Analysis

Encoding is the first step of both frequent itemset mining and association analysis. Both algorithms require that all the data for a transaction (visits) be included in 1 row and the items (clinics) should be one-hot encoded. We need to encode the data into binary data that shows whether a clinic is visited (1) or not (0). “arules” package has its own one-hot encoding function, so encoding is automatically done after listing the clinics next to visitIDs. For other mining tools, you may need to encode data by yourself before feeding the data to the mining tool. A partial list of visitIDs and clinics list for the dataset and one-hot encoded data are given in Figs. 3 and 4. After encoding, a transaction object is created including the one-hot encoded data which has all visit data to all clinics.

In order to find the frequent itemsets, support and confidence levels (*minsupp*, *minconf*) should be selected. Selection of support and confidence does not have a strict rule, instead, different results can be found by changing these levels. In this study, *minsupp* = 0.001 and *minconf* = 0.5 are used by making different trials. Changing these values increases or decreases the number of rules mined from the transaction data. By using these values, thirteen rules are mined from the dataset.

**a**

	patientID	birthyear	gender	submission_date	clinic	insurance
1	22	1971	F	2018-03-29	GYNECOLOGY AND OBSTETRICS	socialsec
2	22	1971	F	2018-04-10	ENDOCRINOLOGY AND METABOLISM DISEASES	socialsec
3	22	1971	F	2018-04-11	NUTRITION AND DIETETIC	cash
4	23	1969	F	2018-06-20	HEMATOLOGY	AK
5	23	1969	F	2018-06-20	DERMATOLOGY	AK
6	23	1969	F	2018-06-25	DERMATOLOGY	AK
7	23	1969	F	2018-07-02	DERMATOLOGY	AK
8	26	1960	M	2018-06-09	DERMATOLOGY	preventiveSS
9	29	2011	M	2018-02-26	CHILD ALLERGY	socialsec
10	35	2010	F	2018-03-08	EYE DISEASES	socialsec

**b**

	patientID	birthyear	gender	submission_date	clinic	insurance	visitID
1	23	1969	F	2018-06-20	HEMATOLOGY	AK	4
2	23	1969	F	2018-06-20	DERMATOLOGY	AK	4
3	37	1963	F	2018-02-12	INTERNAL DISEASES	ss=preventiveSS	11
4	37	1963	F	2018-02-12	GYNECOLOGY AND OBSTETRICS	ss=preventiveSS	11
5	41	1945	M	2018-10-09	NEUROLOGY	socialsec	13
6	41	1945	M	2018-10-09	BIOCHEMISTRY	private	13
7	48	1953	M	2019-05-29	CARDIOLOGY	socialsec	19
8	48	1953	M	2019-05-29	INFECTIOUS DISEASES	socialsec	19
9	58	1960	F	2018-04-17	NEUROLOGY	socialsec	21
10	58	1960	F	2018-04-17	EAR-NOSE-THROAT DISEASES	socialsec	21

**Fig. 2** a and b Original and preprocessed data**Fig. 3** A partial view of the list of clinics for each visitID

visitID	clinic
1260	c("ANESTHESIOLOGY", "UROLOGY")
2800	c("ANESTHESIOLOGY", "SURGERY")
4041	c("ANESTHESIOLOGY", "SURGERY")
7134	c("ANESTHESIOLOGY", "OBSTETRICS AND GYNECOLOGY")
8112	c("ANESTHESIOLOGY", "NEUROSURGERY")
10104	c("ANESTHESIOLOGY", "SURGERY")
11480	c("ANESTHESIOLOGY", "SURGERY")
12649	c("ANESTHESIOLOGY", "OBSTETRICS AND GYNECOLOGY")
12699	c("ANESTHESIOLOGY", "GENERAL PRACTICE")
14416	c("ANESTHESIOLOGY", "SURGERY")
22220	c("ANESTHESIOLOGY", "BIOCHEMISTRY")



	ANESTHESIOLOGY	BIOCHEMISTRY	CARDIOLOGY	CARDIOVASCULAR SURGERY	CHECK UP	CLINICAL ONCOLOGY
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	1	0	0	0	1	0
5	0	0	0	0	0	0
6	0	0	0	0	0	1
7	0	0	0	0	0	0
8	0	1	1	1	0	0
9	0	0	0	0	0	0
10	0	0	0	1	1	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	1	0	1	1	0	0
17	0	0	0	0	0	1
18	0	0	0	0	0	0
19	0	0	0	0	0	1
20	0	0	0	0	0	0

Fig. 4 A partial view of one-hot encoded data

The top-10 most frequently visited clinic sets are given in Fig. 5. Biochemistry, Cardiology, and Ophthalmology are at the top of the list. These are all frequent itemsets with the length of one. A total of 293 itemsets are mined, and 52 of them are used in mining association rules according to chosen support and confidence levels. For association rules mining, any of the three algorithms mentioned in Sect. 3.2 can be utilized. Since the FP-Growth algorithm is not defined in arules package in R, both Eclat and Apriori algorithms are used in association rule mining and gave the same results. They only differ in the method, but the result is the same.

Thirteen association rules which are determined by taking  $minsupp = 0.001$  and  $minconf = 0.5$  are given in Table 4 as sorted by lift values. The list can also be sorted by support, confidence, and coverage.

148 cases are found to have Psychology and Psychiatry together, with support of 0.002 and confidence of 0.908. It means that this rule occurs in 0.2% of all transactions and for 90% of transactions containing psychology, the rule  $\{\{psychology\} \rightarrow$

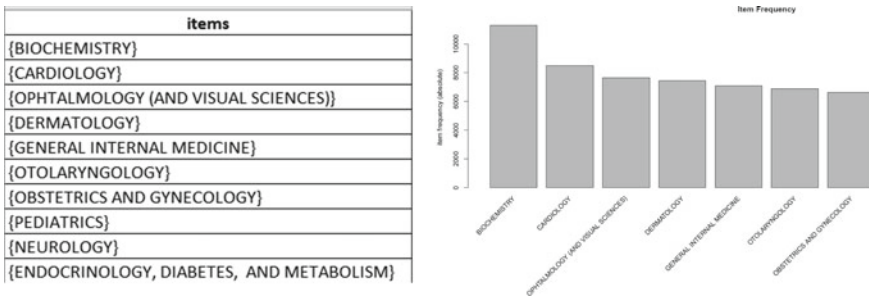


Fig. 5 The top-10 most frequently visited clinic sets

**Table 4** Association rules, sorted by lift

	Rules	Support	Confidence	Coverage	Lift	Count
1	{PSYCHOLOGY} = > {PSYCHIATRY}	0,002	0,908	0,003	<b>18,505</b>	148
2	{CARDIOLOGY,GENERAL INTERNAL MEDICINE,SURGERY} = > {ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY)}	0,004	<b>0,942</b>	0,004	12,943	242
3	{RADIATION ONCOLOGY} = > {CLINICAL ONCOLOGY}	0,006	0,693	0,009	10,462	388
4	{CARDIOLOGY,GENERAL INTERNAL MEDICINE,ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY)} = > {SURGERY}	0,004	0,656	0,006	9,400	242
5	{CARDIOLOGY,ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY),SURGERY} = > {GENERAL INTERNAL MEDICINE}	0,004	0,871	0,004	7,696	242
6	{NUTRITION AND DIETETICS} = > {ENDOCRINOLOGY, DIABETES, AND METABOLISM}	<b>0,008</b>	0,661	<b>0,012</b>	7,396	478
7	{PEDIATRIC SURGERY} = > {PEDIATRICS}	0,007	0,760	0,009	7,379	449
8	{CARDIOLOGY,SURGERY} = > {ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY)}	0,004	0,513	0,009	7,050	278
9	{GENERAL INTERNAL MEDICINE,ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY),SURGERY} = > {CARDIOLOGY}	0,004	0,938	0,004	6,913	242
10	{CARDIOLOGY,ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY)} = > {GENERAL INTERNAL MEDICINE}	0,006	0,594	0,010	5,254	369
11	{ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY),SURGERY} = > {GENERAL INTERNAL MEDICINE}	0,004	0,547	0,008	4,833	258
12	{ORTHOPAEDIC SURGERY(AND TRAUMATOLOGY),SURGERY} = > {CARDIOLOGY}	0,004	0,589	0,008	4,341	278

(continued)

**Table 4** (continued)

	Rules	Support	Confidence	Coverage	Lift	Count
13	{GENERAL INTERNAL MEDICINE, ORTHOPAEDIC SURGERY (AND TRAUMATOLOGY)} = > {CARDIOLOGY}	0,006	0,540	0,011	3,982	369

{psychiatry}) is correct. The lift value of this rule (18,505) is the maximum value and it points out that this rule is the strongest association in all rules.

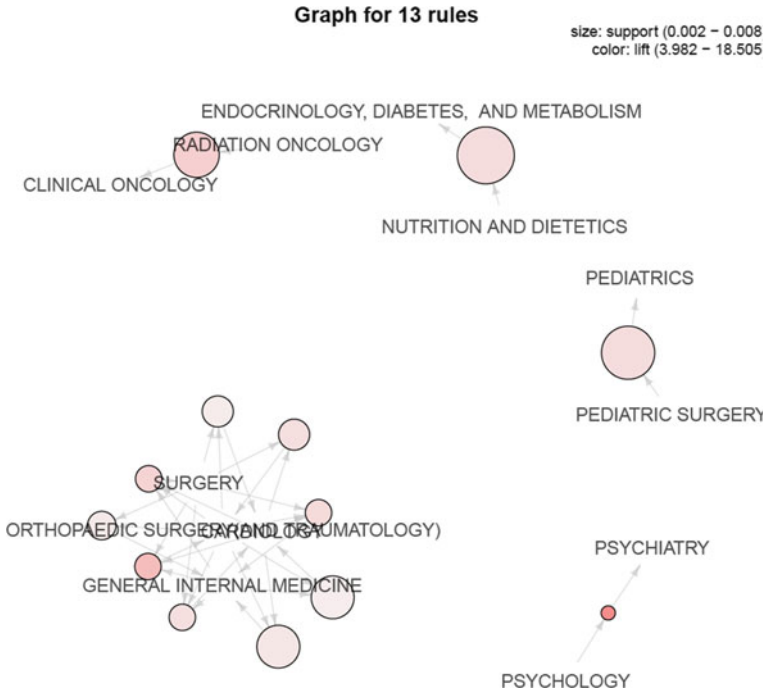
Based on these rules, the maximum support and coverage belong to {nutrition and dietetics} → {endocrinology, diabetes, and metabolism} association. It means that this rule occurs in 0.8% of all transactions and 478 cases have this rule. The highest coverage value (0,012) indicates that this rule is a major pattern. The maximum confidence belongs to {cardiology, general internal medicine, surgery} → {orthopaedic surgery (and traumatology)} association. This rule is correct for 94% of transactions containing cardiology, general internal medicine, and surgery. The rule {pediatric surgery} → {pediatric} is marked in bold because it has the second-highest support value. It means that 0.7% of all transactions have this rule and this corresponds to 449 cases.

This rule list shows the association between clinics. However, some rules are subsets of others. For example, rules 2, 4, 5, 8, 9, 10, 11, 12, and 13 all have the same clinics either on the left or right-hand side. It means that patients visit cardiology, general internal medicine, surgery, and orthopedic surgery clinics together, independent of the order. If we want to group most associated clinics together on the same floor or building, we have to consider this subset list. The subset list reduces the associations to five rules, which are shown in bold in Table 4. A similar conclusion can be reached by examining the rule graph in Fig. 6.

#### 4.4 Visualization of Rules

Mining association rules may result in a very large number of discovered rules, leaving the analyst with the challenge to go through all the rules and summarize the results. Visualization helps in this process. Five criteria are included in the visualization of association rules: sets of antecedent items, consequent items, the association between antecedent and consequent, the support of the rule, the confidence of the rule. Different scatterplots, graphs, and matrix representations may be used for visualization.

Figure 3 shows a “rule graph” where the ruleset is represented with nodes (rules) and arrows (clinics). The size of the nodes shows the support value and the color intensity shows the lift. The five non-overlapping association rules can easily be seen



**Fig. 6** Rule graph

in the figure without making any arrangements. This is the advantage of graphical visualization.

A “matrix visualization” with grouped antecedents for the set of rules may also be drawn as in Fig. 7. The groups of most interesting rules according to lift (the default measure) are shown in the top-left corner of the plot. Since the number of rules is not big in our study, mostly one rule per column is shown. If there were many rules, more than one rule would be shown in each column. In this matrix visualization, confidence is represented by the size of the circle and the color represents the lift.

Order of the associations can be examined in “parallel coordinates plot” where arrow width represents the support and the color intensity represents the confidence. Figure 8 displays the parallel coordinates plot with reordered clinics to minimize crossovers. The tick on the right represents the right-hand side of the rules while the order is represented on the vertical lines beginning from the left of the figure. Here again, we can see visually that cardiology, general internal medicine, surgery, and orthopedic surgery clinics are associated together, mostly with an order of three.

Since our ruleset is small, these visualizations are very useful for interpreting associations. However, if we had a large-volume ruleset, other tools might be necessary. As denoted in arules visualization library (arulesViz) manual (Hahsler and Cheluboina 2019), advanced graphical features such as zooming, filtering, grouping, and coloring nodes are required to explore broad sets of rules with graphs. Such

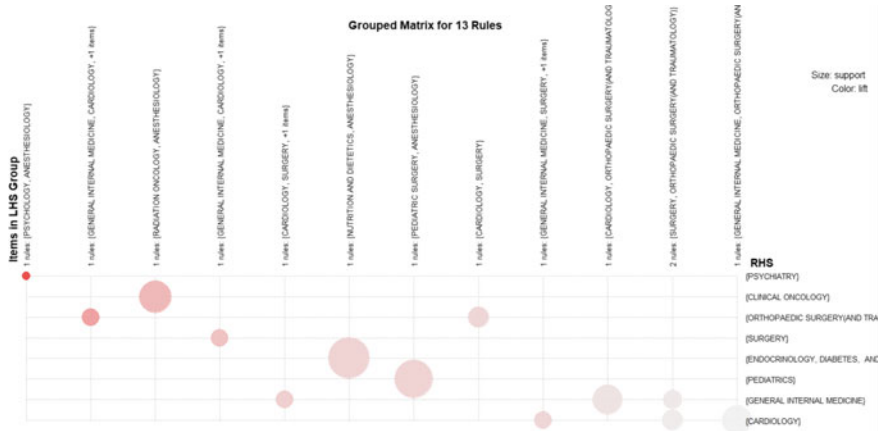


Fig. 7 Matrix visualization of the rules

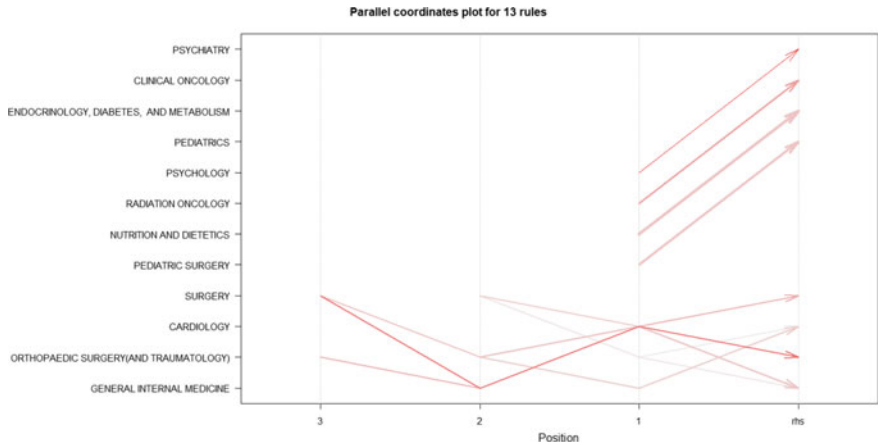


Fig. 8 Parallel coordinates plot

features are included in digital visualization and discovery tools for networks and graphics such as Gephi developed by Bastian et al. (2009).

### 5 Conclusion and Recommendations

This chapter aims to guide the hospital in terms of patient-friendly clinic layouts. The top five most frequently visited clinics are obstetrics and gynecology, cardiology, pediatrics, dermatology, and clinical oncology. These clinics should be as far as possible to avoid patient congestion and crowd. However, cardiology, general internal

medicine, surgery, and orthopedic surgery clinics should be as close as possible since they are all associated with each other. Based on other association rules, pediatrics and pediatric surgery clinics, psychology and psychiatry clinics, nutrition and dietetics and endocrinology clinics, radiation oncology and clinical oncology clinics should be close to each other, too. This brings forward another problem of layout optimization with multiple objectives. Cafes, service areas, toilets, waiting rooms, cashier's desks, information points, phlebotomy rooms, and other support units should be placed according to the results of this association analysis.

Data analytics helps healthcare facilities in many different ways. Association analysis is one of these methods, but unfortunately, it is underrated in health care. Many other areas can be found in health care that the associations play an important role in decision-making. This study shows that it can even be useful in hospital layout and service design.

## References

- Agrawal R, Mehta M, Shafer JC, Srikant R, Arning A, Bollinger T (1996) The quest data mining system. In: KDD, vol 96, pp 244–249
- Badr W (2019) 6 different ways to compensate for missing values in a dataset. <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>. Accessed 26 Feb 2021
- Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks, pp 361–362
- Belyi E, Giabbanelli PJ, Patel I, Balabhadrapathruni NH, Abdallah AB, Hameed W, Mago VK (2016) Combining association rule mining and network analysis for pharmaco-surveillance. *J Supercomput* 72(5):2014–2034
- Birnbaum EBD (2004) Application of data mining techniques to healthcare data. *Infection control and hospital epidemiology*
- Borgelt C (2005) An implementation of the FP-growth algorithm. In: Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, pp 1–5
- Borgelt C (2012) Frequent itemset mining. *Wiley Interdiscip Rev: Data Min Knowl Discov* 2(6):437–456
- Canlas RD (2009) Data mining in healthcare: current applications and issues. School of information systems & management, Carnegie Mellon University, Australia
- Doddi AM, Ravi SS, David C, Torney S (2001) Discovery of association rules in medical data. *Med Inform Internet Med* 26(1):25–33
- Duan L, Street WN, Xu E (2011) Healthcare information systems: data mining methods in the creation of a clinical recommender system. *Enterp Inf Syst* 5(2):169–181
- Goethals B (2003) Survey on frequent pattern mining. *Univ Helsinki* 19:840–852
- Hahsler M (2017) arulesViz: interactive visualization of association rules with R. *R J* 9(2):163–175
- Hahsler M, Chelluboina S (2019) Visualizing association rules: introduction to the r-extension package arulesViz. <https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf>. Accessed 26 Feb 2021
- Hahsler M, Grün B, Hornik K (2005) arules—a computational environment for mining association rules and frequent item sets. *J Stat Softw* 14(15):1–25. <https://doi.org/10.18637/jss.v014.i15>
- Hande R, Bulchandani V, Batreja H, Jaisinghani K, Nagwan S (2015) Mining medical data for identifying frequently occurring diseases by using apriori algorithm. *Int J Comput Appl* 975:8887

- Harahap M, Husein AM, Aisyah S, Lubis FR, Wijaya BA (2018) Mining association rule based on the disease population for recommendation of medicine need. In: *Journal of Physics: Conference Series*, vol 1007, no 1. IOP Publishing, p 012017
- Heaton J (2016) Comparing dataset characteristics that favor the apriori, Eclat or FP-Growth frequent itemset mining algorithms. In: *SoutheastCon 2016*. IEEE, pp 1–7
- Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. *ACM SIGKDD Explor Newsl* 2(1):58–64
- Hornik K, Grün B, Hahsler M (2005) arules—A computational environment for mining association rules and frequent item sets. *J Stat Softw* 14(15):1–25
- Huang F, Wang S, Chan CC (2012) Predicting disease by using data mining based on healthcare information system. In: *2012 IEEE international conference on granular computing*. IEEE, pp 191–194
- Huang YC (2013) Mining association rules between abnormal health examination results and outpatient medical records. *Health Inf Manag J* 42(2):23–30
- Ilayaraja M, Meyyappan T (2013) Mining medical data to identify frequent diseases using apriori algorithm. In: *2013 international conference on pattern recognition, informatics and mobile engineering*. IEEE, pp 194–199
- Jothi N, Husain W (2015) Data mining in healthcare—a review. *Procedia Comput Sci* 72:306–313
- Kaur H, Wasan SK (2006) Empirical study on applications of data mining techniques in healthcare. *J Comput Sci* 2(2):194–200
- Khader N, Lashier A, Yoon SW (2016) Pharmacy robotic dispensing and planogram analysis using association rule mining with prescription data. *Expert Syst Appl* 57:296–310
- Kraft MR, Desouza KC, Androwich I (2003) Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population. In: *Proceedings of the 36th annual hawaii international conference on system sciences, 2003*. IEEE, pp 9–pp
- Lin SL, Wang CS, Chiu HC, Juan CJ (2016) Analyzing medical transaction data by using association rule mining with multiple minimum supports. In: *Pacific Asia Conference On Information Systems (PACIS)*. Association For Information System
- Liu Y (2010) Study on application of apriori algorithm in data mining. In: *2010 Second international conference on computer modeling and simulation*, vol. 3. IEEE, pp 111–114
- Mahindrakar P, Hanumanthappa M (2013) Data mining in healthcare: a survey of techniques and algorithms with its limitations and challenges. *Int J Eng Res Appl* 3(6):937–941
- Mohapatra RK, Pintilie L, Kandi V, Sarangi AK, Das D, Sahu R, Perekhoda L (2020) The recent challenges of highly contagious COVID-19, causing respiratory infections: symptoms, diagnosis, transmission, possible vaccines, animal models, and immunotherapy. *Chem Biol Drug Des* 96(5):1187–1208
- Nahar J, Imam T, Tickle KS, Chen YPP (2013) Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Syst Appl* 40(4):1086–1093
- Ogundele IO, Popoola OL, Oyesola OO, Orija KT (2018) A review on data mining in healthcare. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*
- Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F (2015) An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database. *Int J Endocrinol Metabol* 13(2)
- Rashid MA, Hoque MT, Sattar A (2014) Association rules mining based clinical observations. [arXiv:1401.2571](https://arxiv.org/abs/1401.2571)
- Reps JM, Aickelin, U, Ma J, Zhang Y (2014) Refining adverse drug reactions using association rule mining for electronic healthcare data. In: *2014 IEEE international conference on data mining workshop*. IEEE, pp 763–770
- Saxena A, Gadhya S (2014) A Survey on frequent pattern mining methods—apriori, Eclat, FP growth. *Int J Eng Dev Res* 2(1):92–96
- Schmidt-Thieme L (2004) Algorithmic features of Eclat. In: *FIMI*
- SCI2S (2021) Missing values in data mining. <https://sci2s.ugr.es/MVDM>. Accessed 26 Jan 2021

- Shahin A, Moudani W, Chakik F, Khalil M (2014) Data mining in healthcare information systems: case studies in Northern Lebanon. In: The 3rd International Conference on e-Technologies and Networks for Development (ICeND2014). IEEE, pp 151–155
- Shih WF, Lin CW, Wang WF, Wu HH (2018) Association rule mining of care targets from hospitalized dementia patients from a medical center in Taiwan. *J Stat Manag Syst* 21(7):1299–1310
- Swathi P, Prajna B (2016) The effective procession of apriori algorithm prescribed data mining on medical data. *IJCST* 7(3)
- Țăranu I (2016) Data mining in healthcare: decision making and precision. *Database Syst J* 6(4):33–40
- Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L (2012) Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst* 36(4):2431–2448
- Zhang C, Zhang S (2003). Association rule mining: models and algorithms, vol 2307. Springer, Berlin
- Zhang W, Liao H, Zhao N (2008) Research on the FP growth algorithm about association rule mining. In: 2008 international seminar on business and information management, vol 1. IEEE, pp 315–318
- Zielosko B (2016) Application of dynamic programming approach to optimization of association rules relative to coverage and length. *Fund Inform* 148(1–2):87–105



# Sustainability Transition Through Awareness to Promote Environmental Efficiency



Nikos Chatzistamoulou and Phoebe Koundouri

**Abstract** The 17 Sustainable Development Goals, United Nations' Agenda 2030, the Paris Agreement, the European Green Deal, and the current global policy momentum towards green efficiency, motivate the need for a better understanding of the determinants of environmental efficiency to tackle climate change. By adopting a non-parametric metafrontier framework, the productive performance and environmental efficiency through the Data Envelopment Analysis and Directional Distance Function for each of the 104 countries from 2006 through 2014 are calculated. We contribute to the understanding of environmental efficiency patterns through partitioning the metafrontier via a factor encapsulating 56 environmental indicators to give rise to heterogeneous environmental awareness regimes. By adopting fractional probit models, we show econometrically that productive performance appears to be a major driver of environmental efficiency *only* for the environmentally aware country economies whereas a direct rebound effect is also documented. This is a result with major “policy sequencing” implications. Absorptive capacity reflecting the ability and potentiality of the country to benefit from technological developments seems to play a crucial role as well. The less environmentally aware cluster does not seem to respond the same way to the set of factors considered, indicating that complexity and latent mechanisms affect green efficiency.

**Keywords** Environmental efficiency · Productive performance · Spillover effects · Directional distance function · Sustainability · Green efficiency

---

N. Chatzistamoulou (✉) · P. Koundouri

School of Economics and Research Laboratory On Socio-Economic and Environmental Sustainability–ReSEES, Athens University of Economics and Business, Athens, Greece  
e-mail: [chatzist@aub.gr](mailto:chatzist@aub.gr)

P. Koundouri

e-mail: [pkoundouri@aub.gr](mailto:pkoundouri@aub.gr)

N. Chatzistamoulou

Department of Economics, University of Ioannina, Ioannina, Greece

P. Koundouri

Director, Sustainable Development Unit, ATHENA Research Center, Co-Chair, UN SDSN Europe Fellow, Academy of Art and Science, Marousi, Greece

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

345

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_21](https://doi.org/10.1007/978-3-030-85254-2_21)

# 1 Introduction

Environmental performance enhancement has always been in the center of attention and one of the main pillars of the prosperity at a universal level. Technological heterogeneity, the ability of each country to adopt and internalize technical progress, i.e., absorptive capacity (Cohen and Levinthal, 1990; Zhang et al., 2010), knowledge spillover effects as the carriers of new but potential complex technological achievements and developments that affect performance (Girma, 2005; Casu et al., 2016; Tsekouras et al., 2016) along with the policy directives, all have their own merit on boosting environmental performance. Augmenting the latter argument, under the Sustainable Development Goals Initiative (United Nations, 2015) and the new growth strategy of Europe, that is the European Green Deal<sup>1</sup> (EGD) (European Commission COM (2019) 640), there is a systematic mobilization towards sustainability transition and green growth.

In particular, the Sustainable Development Goals (SDGs) Initiative expressed as targets to be achieved allocated in 17 goals, recently have been restructured in six transformations (Sachs et al., 2019). Those transformations aim at promoting environmental quality through eco-friendly technologies and sustainable ways of production and consumption, among others. Although agreed by the member states, the goals do not constitute an obligation. In this line, the European Green Deal among its main policy areas,<sup>2</sup> includes a climate package for stakeholder engagement referring to every aggregation level (e.g., policy makers, financial institutions, businesses, civil Non-Governmental Organizations, countries) in order to promote commitment and implementation of the directives. The common objectives of the SDGs and the EGD are highlighted via a thorough mapping of the recent report of Koundouri and Sachs (2021) for the Sustainable Development Solutions Network (SDSN, 2021).

It therefore becomes apparent that heterogeneous environmental awareness levels exist across the globe as countries face uneven technological opportunities and access to resources affecting environmental performance. Even though previous studies have employed different factors such as income level and geographical location (e.g., Oh and Lee, 2010) to study performance change, no systematic attempt has been surfaced yet, neither to group countries based on indicators related to the SDGs nor to study environmental awareness through a metaproduction-metafrontier framework. Therefore, it remains a void to be filled.

The contribution of this study is multi-fold. *We adopt the metafrontier-metaproduction framework to account for technological heterogeneity, (ii) we give rise to heterogeneous environmental awareness regimes via partitioning the overall technology by a factor encapsulating aspects of several sustainable development*

---

<sup>1</sup> For a detailed presentation, [https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC\\_1&format=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF) and [https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal\\_en](https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en).

<sup>2</sup> The policy areas included by the EGD are clean energy, climate action, sustainable industry, eliminating pollution, biodiversity, from farm to fork, sustainable agriculture, sustainable mobility, building and renovating.

*goals proxied by environmental indicators, (iii) we study whether the heterogeneous environmental awareness regimes exert differential effect on environmental efficiency of the country economies by employing fractional probit models, and (iv) the technical analysis provided herein contributes to one of the most important sides of environmental transition, that is the stakeholders.*

Findings indicate that awareness matters towards the transition to sustainability. In particular, productive performance appears to be a driver of environmental efficiency *only* for the environmentally aware country economies. Those countries need to further promote and develop the partnerships among the SDGs as well as endorse the directives described in the action plans of EGD. Absorptive capacity seems to play a crucial role as well. A rebound effect is also observed for the global technology as well as for the environmentally aware country economies. However, the less environmentally aware country economies do not seem to respond the same way to the set of factors considered, indicating that complexity and latent mechanisms affect green efficiency. In those countries, the directives regarding stakeholder engagement become predominant, as that would set the latter in a resilient and sustainable trajectory.

This chapter unfolds as follows. The next section offers a brief overview of the related literature, Sect. 3 presents the methodology and research hypotheses, Sect. 4 presents the data, Sect. 5 is dedicated to the results and discussion while Sect. 6 concludes the chapter.

## 2 Related Literature

There is ample literature regarding environmental performance and every attempt to be exhaustive would be unintentionally incomplete. The Porter Hypothesis has been a beacon for research proliferation even though the literature is quite mixed. Indicatively, Rubashkina et al. (2015) test for weak and strong versions of the Porter Hypothesis and relate it to environmental regulation and competitiveness using a panel of manufacturing industries in 17 European Union (EU) countries over the period 1997–2009 to find evidence in favor of the weak version, while productivity appears to be unaffected by the stringency of environmental regulation. Costantini and Crespi (2008) focus on the export flows of environmental technologies across the globe, providing support for the Porter and Van den Linde hypothesis stating that it has brought to the forefront the role of energy policy design as a mechanism towards sustainability. The Kyoto Protocol directives are also in this line boosting innovation in the energy sector. In the same line, Hart (2004) presents theoretical models falling in the context of the endogenous growth theory to model technical change and the environment, concluding that penalizing dirty ways of production is beneficial not only for social utility but also for the improvement of the growth rate of production. Thus, it falls in the group of studies supporting the Hypothesis.

A significant number of studies regarding environmental and energy efficiency, i.e., resource efficiency measures, have surfaced aiming to explore the economy of

China. Chang et al. (2013) analyze the environmental efficiency of China transportation industry by proposing a non-radial Data Envelopment Analysis (DEA) model with slack-based-measures to find that the latter lacks in efficiency. Other sectoral studies, such as the work of Zofio and Prieto (2001) who calculate the environmental efficiency of the Organization for Economic Cooperation and Development (OECD) manufacturing industries under many carbon dioxide emission regulatory scenarios, highlight the use of the non-parametric techniques in assessing environmental performance. Other applications of environmental efficiency estimation include the construction industry in China (Xian et al., 2019) and the international trade and telecommunications industry (Perkins and Neumayer, 2009), just to mention a few. It should be noted that the relationship among environmental policy, environmental performance, and competitiveness depends on the application considered or sector selected (Iraldo et al., 2011).

Cross-nation performance comparisons raise the issue of technological heterogeneity as country economies do not share identical technology and resource endowments affecting performance. Therefore, the need for a methodological framework embracing all possible aspects of heterogeneity is imperative. The concept of the metaproduction function of Hayami (1969) and Hayami and Ruttan (1970) materialized through the metafrontier framework of O'Donnell et al. (2008) set a new perspective in efficiency analysis.

The literature has been expanded to include climate and environmental footprint assessment studies focused on industry applications to explore the effect of sustainable construction on resource efficiency (Tan et al., 2011). Others focus on the environmental tax reform in the EU-27 under the Kyoto protocol, to find that technological spillover effects mitigate the negative effects of carbon leakages (Barker et al., 2007). The impact of spillover effects on resource efficiency measures such as energy efficiency, environmental efficiency, and productive performance, under a technology heterogeneity framework has been acknowledged in a series of recent contributions as well (Tsekouras et al., 2016; Chatzistamoulou et al., 2019).

For instance, Wei et al. (2019) handle heterogeneity by applying the modified method of Metafrontier Malmquist Luenberger Index (MML). They partition the overall technology of the 97 Paris Agreement contracting countries by income level for the period 1990–2014 to find that heterogeneity affects the MML patterns across the groups. Wang et al. (2019) use a variant of the MML on the G20 countries from 2000 to 2014 to make environmental efficiency comparisons as well. Feng and Wang (2019) find positive evidence related to pollution migration in China for the period 2001–2016 as the emissions efficiency improved.

It is therefore evident that despite the empirical studies scattered in the literature, there is a void to be filled regarding the impact of environmental awareness on the environmental efficiency. This is particularly relevant nowadays under the urgency to set economies into a smooth transition trajectory leading to a sustainable future as promoted by the SDGs as well as the EGD.

### 3 Methodological and Theoretical Considerations

#### 3.1 Constructing Environmental Awareness Clusters

To handle technological heterogeneity on the benchmarking process (Dosi et al., 2010), statistical techniques often used to create relatively homogeneous groups in line with the literature (Chui et al., 2012; Lin et al., 2013; Zhang et al., 2014; Wang et al., 2016). We employ the principal component analysis with varimax rotation (Genious et al., 2014), to construct a partitioning factor for the global technology by considering 56 environment indicators mirroring aspects of several sustainable development goals from World Bank Environment Indicators database. Then, we apply the *k-means* clustering procedure to construct two clusters reflecting differences in environmental awareness.

That being said, we construct the environmentally aware (EA) and the less environmentally aware (LEA) cluster, respectively. In this context, *an environmental awareness regime is considered as a production frontier to benchmark the country economies. It encompasses the technological complexity, differences in resource endowments, country objectives to preserve environmental quality, and efforts to mitigate negative effects of climate change through implementing an active strategy of protecting scarce resources.* This paves the way to investigate the effect of a plethora of technological and environmental aspects on environmental efficiency on a global scale to promote sustainability.

Environmental awareness as means to proxy the mindset towards sustainability of production gains ground gradually. That being said, Giudici et al. (2019) investigate the effect of sensitivity to environmental issues by local governments, firms, and residents, framing the former as local environmental awareness on green start-ups creation. Although it is feasible to create more than two groups, the number of entities under each environmental awareness production frontier would be reduced and more entities would have been falsely identified as fully efficient (Dyson et al., 2001).

#### 3.2 Performance Evaluation Under Heterogeneity

##### 3.2.1 Productive Performance; The Data Envelopment Analysis Technique

A country economy  $i = 1, 2, \dots, n$  may be considered as a production entity transforming inputs  $x = (x_{1i}, x_{2i}, \dots, x_{Ni},) \in \mathfrak{R}_+^N$  into outputs  $y = (y_{1i}, y_{2i}, \dots, y_{Mi},) \in \mathfrak{R}_+^M$  under a technology set  $S$  defined as  $S \equiv \{(x, y) : x \text{ can produce } y\}$ . For the input-oriented productive performance, the technology is represented by the production possibility set  $L(y) = \{x \in \mathfrak{R}_+^N : (x, y) \in S\}$ , while for its measurement the input distance function defined as  $D_I(x, y) = \sup\{\theta > 0 : x/\theta \in L(y)\}$  is used. In the case where

two environmental awareness production frontiers (technologies)  $T^{EA}, T^{LEA}$  exist, the metatechnology set, denoted as  $T^M$ , can be defined as the convex hull of the jointure of the two technology sets represented as  $T^M = \{(x, y : x \geq 0, y \geq 0) \text{ can produce at least one of } T^{EA}, T^{LEA}\}$  (Battese et al., 2004). The technology set can be defined in the same way for the single technology.

By adopting the metafrontier framework (global technology-MF) as introduced by Hayami (1969) and Hayami and Ruttan (1970) and further developed by O’Donnell et al. (2008), and employing the bootstrap version of the input-oriented DEA under variable returns to scale to account for size effects (Halkos and Tzeremes, 2009), the bias corrected productive performance of each country economy with respect to the global technology is calculated using the following formula:

$$\widehat{ProdPerf}_{i,t}^{MF} \equiv \hat{\theta}(x, y) = \min \left\{ \theta \mid \theta > 0, y \leq \sum_{i=1}^n \gamma_i y_i; \theta x \geq \sum_{i=1}^n \gamma_i x_i \text{ for } \gamma_i \right. \quad (1)$$

such that

$$\left. \sum_{i=1}^n \gamma_i = 1; \gamma_i \geq 0, i = 1, 2, \dots, K \right\}$$

Productive performance (*ProdPerf*) of each country economy is calculated within each cluster by employing Eq. 1. The metatechnology ratio (*MTR*) and the corresponding technology gap (*Tg*) are calculated for each country economy on an annual basis, using the formulas below:

$$MTR_{i,t}(x, y) = \frac{\widehat{ProdPerf}_{i,t}^{MF}(x, y)}{\widehat{ProdPerf}_{i,t}(x, y)} \quad (2)$$

$$Tg_{i,t}(x, y) = 1 - MTR_{i,t}(x, y) \quad (3)$$

The technology gap measures the distance between the individual frontier and the metafrontier capturing spillover effects (Chatzistamoulou et al., 2019).

### 3.2.2 Environmental Efficiency; The Directional Distance Functions Approach

Assuming that the production technology  $T$  models the transformation of a vector of inputs  $x \in \mathfrak{R}_+^N$  that each country economy employs to produce a vector of outputs  $y^* \in \mathfrak{R}_+^M$  as presented in the work of Chambers et al. (1996), Chung et al. (1997), and Fare and Grosskopf (2000), we discern the desirable  $y = (y_1, y_2, \dots, y_k) \in \mathfrak{R}_+^K$  and

the undesirable output  $b = (b_1, b_2 \dots, b_l) \in \mathfrak{R}_+^L$ , respectively<sup>3</sup> (Kumar and Khanna, 2009). The underlying production process is constrained by the technology set (Chambers et al., 1996; Kumar, 2006; Luenberger, 1992; 1995; Shepard, 1953; 1970; Zhou et al., 2012)  $T$  defined as  $T(x) = \{(y, b) : x \text{ can produce } (y, b)\}$  (Dervaux et al., 2009). The directional distance function (DDF) may be represented by a multi-input and multi-output distance function on technology  $T$  (e.g., Chambers et al., 1998; Picazo-Tadeo et al., 2011) and can be defined as:

$$\overrightarrow{D}_T(x, y, b; g_y, g_b) = \max\{\beta^* : (x, y + \beta^* g_y, b - \beta^* g_b) \in T(x, y, b)\} \quad (4)$$

The input–output vector  $(x, y)$  is projected onto the technology frontier in the  $(g_y, -g_b)$  direction which allows desirable outputs to be proportionally increased, whereas undesired output(s) to be proportionally decreased. The maximum attainable expansion of desirable outputs in direction  $(g_y)$  and the largest feasible contraction of the undesirable outputs in direction  $(-g_b)$  is of major interest. Since the technology set is restricted only to the production of desired output, the environmental efficiency at the metafrontier,  $EnvEff^{MF}$ , is:

$$EnvEff^{MF} = \frac{\left(1 + \overrightarrow{D}_T^{MF}(x, y, b; g_y, g_b)\right)}{\left(1 + \overrightarrow{D}_T^{MF}(x, y, b; g_y)\right)}, \quad (5)$$

with the environmental efficiency for the individual environmental awareness frontiers to be defined in an analogous manner.

The environmental efficiency index ( $EnvEff^{MF}$ ) captures the contraction in increasing outputs by each country economy under the potential ability of the production process convention from free disposability to costly disposal of carbon dioxide emissions taking values between zero and one. For a DMU with environmental efficiency score of one, the cost of transforming their production from strong disposability to weak for emissions should be zero while moving to the opposite direction is considerably costly (Kounetas and Zervopoulos, 2019; Kumar and Khanna, 2009). Environmental efficiency has been defined as the ratio of two distance functions assuming strong and weak disposability of the undesired output, however, the ratio of those two distances leads to values very close or equal to one due to the weak disposability assumption (Zaim and Taskin, 2000).

---

<sup>3</sup> Note that the two different output sets are actually sub-vectors of the  $y^* \in \mathfrak{R}_+^M$  output set.

### 3.3 *Econometric Strategy and Research Questions*

#### 3.3.1 **Fractional Regression Models**

The second stage analysis following the DEA during the past decades, employs mostly binary response dependent variable models to explain the variability in the performance scores attained by the first stage (Gillen and Lall, 1997; Merkert and Hensher, 2011). A systematic review of modelling second stage DEA scores is provided by Hoff (2007).

Papke and Wooldridge (1996; 2008) introduce a more appropriate methodology to handle variables that come in proportions, shares, and in general variables that vary between zero and one. Particularly, in the case of efficiency scores, despite the popular use of limited dependent and censored variable models those (i) cannot adequately capture the nature of the variable, (ii) the censoring does not appear to be applicable if the variable of interest does not exceed those boundaries, (iii) the mechanics of linear models are not capable in handling incremental changes of the explanatory variables on the dependent especially as the latter crowd closely at the boundaries making inappropriate to predict the expected values at the corners (Noreen, 1988; Maddala, 1991; Papke and Wooldridge, 1996; Gallani et al., 2015).

To cope with the limitations of the abovementioned models, Papke and Wooldridge (1996; 2008) propose and develop the idea of fractional regression models (FRM) without the requirement of data transformation at the tails whereas Greene (2003) mentions that FRM exploit data non-linearities to calculate the average partial effects at different percentiles of the predictor(s) distribution. Criticism on the FRM is found on the grounds that the latter do not apply to repeated measurements but since we consider each year as a separate production function, it is consistent with our approach. Structural parameters are estimated via quasi-maximum likelihood which produces robust and relatively efficient estimates, under the GLM assumptions (Gallani et al., 2015). All in all, since environmental efficiency range between zero and one should be considered as fraction and that is the reason why we exploit its potential herein.

#### 3.3.2 **Modelling Environmental Efficiency**

We specify and estimate the following models for the global technology level as well as for the two environmental awareness clusters, by employing three pooled fractional probit models:

$$\begin{aligned}
 EnvEff_{it}^{Global\ technology} &= \beta_0 + \beta_1 ProdPerf_{it} + \beta_2 Spillovers_{it-1} \\
 &+ \beta_3 AC_{it-1} + \beta_4 FraserIndex_{it} \\
 &+ \beta_6 EconStruIndex_{it} + \beta_7 Renewables_{it} \\
 &+ \beta_8 Switch_{it} + \gamma YearEffects + u_{it}
 \end{aligned} \tag{6}$$



$$\begin{aligned}
 EnvEff_{it}^{EA} &= \delta_0 + \delta_1 ProdPerf_{it} \\
 &+ \delta_2 Spillovers_{it-1} + \delta_3 AC_{it-1} \\
 &+ \delta_4 FraserIndex_{it} + \delta_6 EconStruIndex_{it} \\
 &+ \delta_7 Renewables_{it} + \delta_8 Switch_{it} \\
 &+ \rho YearEffects + v_{it}
 \end{aligned} \tag{7}$$

$$\begin{aligned}
 EnvEff_{it}^{LEA} &= \lambda_0 + \lambda_1 ProdPerf_{it} \\
 &+ \lambda_2 Spillovers_{it-1} + \lambda_3 AC_{it-1} \\
 &+ \lambda_4 FraserIndex_{it} + \lambda_6 EconStruIndex_{it} \\
 &+ \lambda_7 Renewables_{it} + \lambda_8 Switch_{it} \\
 &+ \rho YearEffects + v_{it}
 \end{aligned} \tag{8}$$

where  $EnvEff_{it}^{Globaltechnology}$ ,  $EnvEff_{it}^{EA}$ , and  $EnvEff_{it}^{LEA}$  correspond to the environmental efficiency of the  $i$ -country economy in year  $t$  with respect to the global technology as well as of each cluster considered.

We formulate and test three research hypotheses corresponding to each variable of interest, the productive performance ( $ProdPerf_{it}$ ), absorptive capacity ( $AC_{it-1}$ ) captured by the lagged value of competitiveness level, and spillovers ( $Spillovers_{it-1}$ ) captured by the lagged value of technology gap as drivers of the environmental efficiency. Particularly:

H<sub>1</sub>: Productive performance exerts a positive and significant influence on environmental efficiency.

The role of absorptive capacity has been acknowledged by the literature indicating the ability to transform technological achievements into improved performance (Cohen and Levinthal, 1989; 1990; Eichhammer and Walz, 2011). The lagged global competitiveness index (GCI) which is country-specific and time-varying has been used to capture a country's absorptive capacity (Gkypali et al., 2019) while it reinforces the ability and potentiality to absorb accumulated knowledge generated universally. In the form of a testable hypothesis, it can be stated as:

H<sub>2</sub>: Absorptive capacity enhances environmental efficiency.

By rejecting the null would imply that low technological opportunities and assimilation ability negatively affect the environmental efficiency. The influence of spillovers in explaining performance patterns has been acknowledged before (Tsekouras et al. 2016; Chatzistamoulou et al. 2019), thus it is reasonable to include it in explaining performance patterns. This can be stated as:

H<sub>3</sub>: Spillover effects exert a positive and significant influence on the environmental efficiency of each cluster.

Additional variables such as the Frazer index ( $FraserIndex_{it}$ ) and the Economy Structure index  $EconStrIndex_{it}$ <sup>4</sup> have been included to capture characteristics of the overall production environment of each country economy.  $Rec_{it}$  is the share of renewable energy consumption capturing the use of resource-saving paving the way for environmental efficiency improvement. The variable  $Switch_{it}$  captures switches between the two clusters at the global technology level. Year effects have been included while  $u_{it}$ ,  $v_{it}$ , and  $v_{it}$  are the disturbance terms. The parameters to be estimated are  $\beta$ ,  $\delta$ ,  $\lambda$ ,  $\gamma$ ,  $\rho$ , and  $\varrho$ .

## 4 Data

We devise a unique panel by coordinating, matching, and harmonizing several distinct yet complementary publicly available databases covering 104 country economies over a nine-year period, from 2006 through 2014. Therefore, the panel includes 936 observations.

The novelty of this dataset is found on the use of 56 indicators referring to the use of natural resources, changes in the natural and built environment encompassing the availability and use of environmental resources related to environmental degradation, in creating the partitioning factor to give rise to alternative environmental awareness clusters. The indicators mirror and illuminate many aspects of a wide variety of the SDGs such as 2, 6, 7, 11, 12, 14, and 15 (World Bank, 2018). It is not worthless to mention that this is the first time such data are employed to explore environmental awareness.

We collect data on two outputs and three inputs. Outputs include the Gross Domestic Product (GDP) capturing the desired output (measured in mil. US\$) and the carbon dioxide emissions ( $CO_2$ ) capturing the undesired output (measured in kt). Inputs include the capital captured by the capital stock (measured in mil. US\$), labor proxied by the number of persons engaged (measured in mil.), and the energy captured by the energy use (measured in kt of oil equivalent) of each country economy, respectively. Monetary values are in constant 2011 prices.

Additional variables have been collected to account for as many as possible aspects of the production environment. Particularly, absorptive capacity (Cohen and Levinthal, 1990) of each country economy captured by the lagged GCI encapsulating 12 pillars representing each country's potential and market conditions among others (Sala-i-Martin and Artadi, 2004; Sala-i-Martin et al., 2008) produced annually by the World Economic Forum, has proved a quite useful tool in the empirical analysis (Tsekouras et al., 2016; 2017; Chatzistamoulou et al., 2019, Gkypali et al., 2019). The structure of the economy, proxied by the contribution of the secondary, manufacturing, services sectors, the share of renewable energy use to the total energy use,

---

<sup>4</sup> The economy structure index which has been created by combining the share of secondary, manufacturing and services sector on the national product.

and data on the economic freedom captured by the multi-faceted Fraser index have been included.

Data on the Gross Domestic Product, Labor, and Capital have been collected through the Groningen Growth and Development Centre (GGDC), World Penn Tables 9.0. Data on the Environment indicators have been collected through the World Sustainable Indicators (WSI) database of the World Bank. Carbon dioxide emissions, energy use, renewable energy use, secondary, manufacturing and services sector contribution to the gross domestic product have been collected through the World Bank. Data on the GCI has been hand-collected through various releases of the Global Competitiveness Report published by the World Economic Forum annually, while data on the Economic Freedom index was collected through the Fraser Institute official site. Table 1 below provides the basic descriptive statistics and source of the main variables.

## 5 Results and Discussion

Table 2 below presents the estimation results (marginal effects) of the fractional probit models in Eqs. 6–8. The first column corresponds to the estimation results for the case of the global technology. Productive performance at the global level does not seem to be a driver of environmental efficiency ( $H_1$  is not accepted). This is in line with the study of Chatzistamoulou et al. (2019), who consider another resource efficiency measure that of the energy efficiency, to find that productive performance at the global level does not appear to be one of its drivers.

Absorptive capacity seems to exert a positive and significant influence on environmental efficiency as it captures the ability to internalize and exploit any technological and institutional opportunity to enhance performance (Cohen and Levinthal, 1989; 1990). Taking the latter into consideration, under the borderless technology, every country economy has the potential to be benefited by the existence of technological achievements. Even though the assimilation ability and internalization mechanisms may not be similar, it seems that a positive effect arises ( $H_2$  is not rejected). Although many proxies of competitiveness are potentially available (Balkyte and Tvaronavičienė, 2010), those are characterized by subjectivity, as only one aspect is being considered. The multi-faceted GCI accommodates for several pillars<sup>5</sup> common across country economies facilitating comparisons.

The conditions of the production environment appear to be a significant driver in explaining environmental efficiency. Specifically the economy structure index indicates that if the composition of the production environment at the country level has not incorporated clean technologies, negatively affects resource efficiency (York

---

<sup>5</sup> Pillars include Institutions, Infrastructure, Macroeconomic Environment, Health and Primary Education, Higher Education and Training, Goods market efficiency, Financial market development, technological readiness, market size, business sophistication and innovation.

**Table 1** Basic descriptive statistics (Means and St. Dev.) for the main variables, 2006–2014

	Brief description	Units of measurement	Source	All countries - Global Technology	Environmentally Aware	Less Environmentally Aware
<i>GDP</i>	Real Gross Domestic Product	million US \$	GGDC	825,045 (2,129,040)	903,698 (2,254,718)	736,381 (1,976,624)
<i>CO<sub>2</sub></i>	Carbon dioxide emissions	kiloton (kt)	World Bank	287,336 (1,026,434)	324,622 (1,122,714)	245,304 (905,157)
<i>K</i>	Capital stock	million US \$	GGDC	2,855,233 (7,207,575)	3,131,751 (7,705,912)	2,543,522 (6,595,734)
<i>L</i>	Persons engaged	millions	GGDC	26,894 (90,768)	28,890 (99,073)	24,644 (80,437)
<i>E</i>	Energy use	kg of oil equivalent per capita	World Bank	2,363 (2,302)	2,592 (2,384)	2,105 (2,180)

**Table 2** Estimation results—marginal effects

	Global technology	Environmentally Aware cluster	Less Environmentally Aware cluster
<i>Performance measures</i>			
Productive performance	0.010 (0.008)	− 0.053* (0.029)	0.041 (0.036)
Spillovers	−	− 0.037 (0.024)	0.087 (0.056)
Absorptive capacity	0.002** (0.001)	0.006 (0.004)	0.002 0.003
<i>Aspects of production environment</i>			
Economy structure index	− 0.002* (0.001)	− 0.002 (0.001)	− 0.001 (0.002)
Frazer index	0.000 (0.001)	0.001 (0.001)	0.003 (0.003)
Renewables	− 0.000 + ** (0.000)	− 0.000 + * (0.000)	− 0.000 + (0.000)
Regime switches	0.001 (0.002)	−	−
Year effects	Yes	Yes	Yes
<i>Model information</i>			
Log-likelihood	− 15.826	− 6.939	− 11.382
Obs	760	375	370
Model p-value	0.000	0.026	0.004

Notes (i) all models include constants, (ii) robust standard errors in parentheses, (iii) stars indicate statistical significance at 1%\*\*\*, 5% \*\*, 10% \*, (iv) “+” indicates a very small number

et al., 2003; Carattini et al., 2015). This is not the case for the Fraser index which seems to have influence on environmental efficiency.

However, there is a negative influence triggered by increased use of renewables which pinpoints towards a direct rebound effect (Binswanger, 2001; Hertwich, 2005; Deng and Newton, 2017). Last but not least, given that cluster switching does not systematically affect environmental efficiency, indicates that production paradigms take time to change, and country economies need time to adjust, internalize, and reform.

Shifting the attention to the environmentally aware cluster, we find that productive performance exerts a negative but significant influence on environmental efficiency ( $H_1$  is partially accepted). This finding indicates that the two performance measures are not heading towards the same direction. This finding comes with policy sequencing implications when designing environmental policies to promote performance. This could be facilitated by the introduction of a more concrete legal framework that provides the incentive to replace existing technologies with one that are

more environmentally attuned so as to develop greener production scenarios. This is not a peculiar finding as a similar relationship between productive performance and energy efficiency has been documented before (Chatzistamoulou et al., 2019). Absorptive capacity does not exert a systematic effect on environmental efficiency ( $H_2$  is not accepted), indicating that in order to promote assimilation of technological achievements, pillars should be improved (Sala-i-Martin et al., 2008). Country economies of this cluster appear to have limited potential to accommodate technological achievements or opportunities for catch-up with the current developments ( $H_3$  is not accepted). The latter could be attributed to the localized nature of spillovers, as economy sectors are not equally developed across countries. In this line, Braun et al. (2010) highlight the distinct nature of spillovers, to those from the same and other related technologies. Furthermore, the systematic effect of the use of renewables on environmental efficiency, pinpoints towards a rebound effect.

Finally, focusing on the less environmentally aware cluster, it is evident that there is a great deal of complexity. The factors affecting environmental efficiency of the environmental aware cluster have a differential effect in this case ( $H_1$ – $H_3$  are not accepted). Such finding underlines the necessity to take technological heterogeneity into account. However, the fact that spillovers appear to exert a rather weak effect on environmental efficiency of this cluster, could indicate that those country economies do not manage to exploit knowledge and technological achievements due to the intrinsic complexity. It is not uncommon for technological knowledge to be tangled and its diffusion proves to be problematic and hard to be assimilated due to the complexity embodied (Kogut and Zander, 1992), especially with environmental practices developed in countries with more advanced technological domains (Rivkin, 2000).

Therefore, a one size-fits-all policy regarding enhancing the environmental performance does not appear to be an appropriate strategy, a tailored set of measures for sophisticated intervention could have an impact instead. Nevertheless, results should be considered with caution as this is the first attempt to study the impact of sustainability, as mirrored by the environment indicators. Results leave the discussion on the drivers of environmental efficiency open for fruitful discussion.

## 6 Conclusions and Remarks

Resource efficiency has been put in the center of the public agenda to lead a smoother transition to sustainability. This has attracted the attention globally, however to a different extent due to the technological, institutional, and other idiosyncratic characteristics of each country economy. It thus becomes apparent that the extent of environmental awareness, protection directives, and guidelines follow heterogeneous patterns universally. This needs to be accommodated in the analysis when attempting to evaluate performance patterns. The efficiency analysis toolbox has been extended to incorporate the Directional Distance Function technique to provide calculations on the environmental efficiency of the production entities to monitor their performance.

To study environmental efficiency under alternative environmental awareness production frontiers, we devise a balanced panel including 104 country economies over a nine-year period, from 2006 to 2014. Then, we employ the non-parametric metafrontier framework and the bootstrap Data Envelopment Analysis under variable returns to scale, to calculate the bias corrected productive performance and technology gap values annually. The environmental efficiency is calculated through the Directional Distance Functions approach. We investigate the drivers of environmental efficiency, through a fractional probit model.

Findings show a quite differentiated mosaic of effects depending on the cluster considered. For the global technology, productive performance does not seem to be the main driver, but this is not the case for absorptive capacity. Productive performance appears to have a significant effect only on the environmentally aware country economies. However, the less environmentally aware cluster does not seem to respond the same way on the drivers explored. The latter might be an indication of technological complexity meaning that knowledge is localized, rigid, and hardly transferable. This highlights the need for restructuring the production paradigm and build on the aspects of the economy that could be used to adopt externally generated knowledge such as a coherent institutional framework, human capital, and technology stock to recombine available resource endowments.

It goes without saying that this study is not free of limitations. First and foremost, more indicators could be considered in order to get a better representation on environmental awareness across countries for a longer period of time to let the effects diffuse to the system, should more data become readily available. Also, policy-related variables could be incorporated in the analysis to explain environmental performance. However, those are latent since there is not an official registry for each country with implementation details, for the time being.

**Acknowledgements** The authors acknowledge the Athens University of Economics and Business Research Centre as the funding source in the context of the Action II Research Support to Post-doctoral Researchers Program 2018–2019 with project code EP-2992-01. The usual disclaimer applies.

## References

- Balkyte A, Tvaronavičiene M (2010) Perception of competitiveness in the context of sustainable development: facets of “sustainable competitiveness.” *J Bus Econ Manag* 11(2):341–365
- Battese GE, Rao DP, O’Donnell CJ (2004) A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies. *J Prod Anal* 21(1):91–103
- Barker T, Junankar S, Pollitt H, Summerton P (2007) Carbon leakage from unilateral environmental tax reforms in Europe, 1995–2005. *Energy Policy* 35(12):6281–6292. <https://doi.org/10.1016/j.enpol.2007.06.021>
- Binswanger M (2001) Technological progress and sustainable development: what about the rebound effect? *Ecol Econ* 36(1):119–132

- Carattini S, Baranzini A, Roca J (2015) Unconventional determinants of greenhouse gas emissions: the role of trust. *Environ Policy Gov* 25(4):243–257
- Casu B, Ferrari A, Girardone C, Wilson JO (2016) Integration, productivity and technological spillovers: evidence for eurozone banking industries. *Eur J Oper Res* 255(3):971–983
- Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. *J Econ Theory* 70(2):407–419
- Chang Y, Zhang N, Danao D, Zhang N (2013) Environmental efficiency analysis of transportation system in China: a non-radial DEA approach. *Energy Policy* 58:277–283. <https://doi.org/10.1016/j.enpol.2013.03.011>
- Chatzistamoulou N, Kounetas K, Tsekouras K (2019) Energy efficiency, productive performance and heterogeneous competitiveness regimes. Does the dichotomy matter? *Energy Econ* 81:687–697
- Cohen WM, Levinthal DA (1989) Innovation and learning: the two faces of R & D. *Econ J* 99(397):569–596
- Cohen WM, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. *Adm Sci Q* 128–152
- Chiu CR, Liou JL, Wu PI, Fang CL (2012) Decomposition of the environmental inefficiency of the meta-frontier with undesirable output. *Energy Econ* 34(5):1392–1399
- Chung YH, Färe R, Grosskopf S (1997) Productivity and undesirable outputs: a directional distance function approach. *J Environ Manag* 51(3):229–240
- Costantini V, Crespi F (2008) Environmental regulation and the export dynamics of energy technologies. *Ecol Econ* 66(2–3):447–460. <https://doi.org/10.1016/j.ecolecon.2007.10.008>
- Deng G, Newton P (2017) Assessing the impact of solar PV on domestic electricity consumption: exploring the prospect of rebound effects. *Energy Policy* 110:313–324
- Dervaux B, Leleu H, Minvielle E, Valdmanis V, Aegerter P, Guidet B (2009) Performance of French intensive care units: a directional distance function approach at the patient level. *Int J Prod Econ* 120(2):585–594
- Dosi G, Lechevalier S, Secchi A (2010) Introduction: Interfirm heterogeneity—nature, sources and consequences for industrial dynamics. *Ind Corp Chang* 19(6):1867–1890
- Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA (2001) Pitfalls and protocols in DEA. *Eur J Oper Res* 132(2):245–259
- Färe R, Grosskopf S (2000) Theory and application of directional distance functions. *J Prod Anal* 13(2):93–103
- Feng C, Wang M (2019) The heterogeneity of China's pathways to economic growth, energy conservation and climate mitigation. *J Clean Prod* 228:594–605. <https://doi.org/10.1016/j.jclepro.2019.04.326>
- Girma S (2005) Absorptive capacity and productivity spillovers from FDI: a threshold regression analysis. *Oxford Bull Econ Stat* 67(3):281–306
- Gallani S, Krishnan R, Wooldridge JM (2015) Applications of fractional response model to the study of bounded dependent variables in accounting research. Harvard Business School
- Genius M, Koundouri P, Nauges C, Tzouvelekas V (2014) Information transmission in irrigation technology adoption and diffusion: social learning, extension services, and spatial effects. *Am J Agr Econ* 96(1):328–344
- Gillen D, Lall A (1997) Developing measures of airport productivity and performance: an application of data envelopment analysis. *Transp Res Part e: Logist Transp Rev* 33(4):261–273
- Giudici G, Guerini M, Rossi-Lamastra C (2019) The creation of cleantech startups at the local level: the role of knowledge availability and environmental awareness. *Small Bus Econ* 52(4):815–830
- Gkypali A, Kounetas K, Tsekouras K (2019) European countries' competitiveness and productive performance evolution: unraveling the complexity in a heterogeneity context. *J Evol Econ* 29(2):665–695
- Halkos GE, Tzeremes NG (2009) Exploring the existence of Kuznets curve in countries' environmental efficiency using DEA window analysis. *Ecol Econ* 68(7):2168–2176



- Hart R (2004) Growth, environment and innovation—a model with production vintages and environmentally oriented research. *J Environ Econ Manag* 48(3):1078–1098. <https://doi.org/10.1016/j.jeem.2004.02.001>
- Hayami Y (1969) Sources of agricultural productivity gap among selected countries. *Am J Agr Econ* 51(3):564–575
- Hayami Y, Ruttan VW (1970) Agricultural productivity differences among countries. *Am Econ Rev* 60(5):895–911
- Hertwich EG (2005) Consumption and the rebound effect: an industrial ecology perspective. *J Ind Ecol* 9(1–2):85–98
- Hoff A (2007) Second stage DEA: comparison of approaches for modelling the DEA score. *Eur J Oper Res* 181(1):425–435
- Iraldo F, Testa F, Melis M, Frey M (2011) A literature review on the links between environmental regulation and competitiveness. *Environ Policy Gov* 21(3):210–222. <https://doi.org/10.1002/eet.568>
- Kogut B, Zander U (1992) Knowledge of the firm, combinative capabilities, and the replication of technology. *Organ Sci* 3(3):383–397
- Kounetas K, Zervopoulos PD (2019) A cross-country evaluation of environmental performance: is there a convergence-divergence pattern in technology gaps? *Eur J Oper Res* 273(3):1136–1148. <https://doi.org/10.1016/j.ejor.2018.09.004>
- Kumar S, Khanna M (2009) Measurement of environmental efficiency and productivity: a cross-country analysis. *Environ Dev Econ* 14(4):473–495
- Kumar S (2006) Environmentally sensitive productivity growth: aglobal analysis using Malmquist-Luenberger index. *Ecol Econ* 56(2):280–293
- Li J, Lin B (2019) The sustainability of remarkable growth in emerging economies. *Resour Conserv Recycl* 145:349–358. <https://doi.org/10.1016/j.resconrec.2019.01.036>
- Lin EYY, Chen PY, Chen CC (2013) Measuring the environmental efficiency of countries: a directional distance function metafrontier approach. *J Environ Manage* 119:134–142
- Maddala GS (1986) Limited-dependent and qualitative variables in econometrics (No. 3). Cambridge University Press
- Maddala GS (1991) A perspective on the use of limited-dependent and qualitative variables models in accounting research. *Account Rev* 66(4):788–807
- Merkert R, Hensher DA (2011) The impact of strategic management and fleet planning on airline efficiency—a random effects Tobit model based on DEA efficiency scores. *Transportation*
- Noreen E (1988) An empirical comparison of probit and OLS regression hypothesis tests. *J Account Res* 119–133
- O’Donnell CJ, Rao DP, Battese GE (2008) Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Empir Econ* 34(2):231–255
- Oh DH, Lee JD (2010) A metafrontier approach for measuring Malmquist productivity index. *Empir Econ* 38(1):47–64
- Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *J Appl Economet* 11(6):619–632
- Papke LE, Wooldridge JM (2008) Panel data methods for fractional response variables with an application to test pass rates. *J Econ* 145(1–2):121–133
- Perkins R, Neumayer E (2009) Transnational linkages and the spillover of environment-efficiency into developing countries. *Glob Environ Chang* 19(3):375–383. <https://doi.org/10.1016/j.gloenvcha.2009.05.003>
- Picazo-Tadeo AJ, Gómez-Limón JA, Reig-Martínez E (2011) Assessing farming eco-efficiency: a data envelopment analysis approach. *J Environ Manage* 92(4):1154–1164
- Rivkin JW (2000) Imitation of complex strategies. *Manage Sci* 46(6):824–844
- Rubashkina Y, Galeotti M, Verdolini E (2015) Environmental regulation and competitiveness: empirical evidence on the porter hypothesis from European manufacturing sectors. *Energy Policy* 83:288–300. <https://doi.org/10.1016/j.enpol.2015.02.014>

- Sachs J, Schmidt-Traub G, Mazzucato M, Messner D, Nakicenovic N, Rockström J (2019) Six transformations to achieve the SDGs. *Nat Sustain*
- Sala-i-Martin X, Blanke J, Hanouz MD, Geiger T, Mia I, Paua F (2008) The global competitiveness index: prioritizing the economic policy agenda. *Glob Compet Rep*, 2009:3–41
- Sala-i-Martin X, Artadi EV (2004) The global competitiveness index. *Glob Compet Rep* 2005:51–80
- SDSN (2021) Transformations for the Joint Implementation of Agenda 2030 for sustainable development and the European green deal. Sustainable development solutions network (SDSN). Lead authors, Koundouri Phoebe and Jeff Sachs
- Shephard RW (1953) Cost and production functions. Princeton University Press, Princeton
- Shephard RW (1970) Theory of cost and production. Princeton University Press, Princeton
- Research part a: policy and practice, 45(7):686–695
- Tan Y, Shen L, Yao H (2011) Sustainable construction practice and contractors' competitiveness: a preliminary study. *Habitat Int* 35(2):225–230. <https://doi.org/10.1016/j.habitatint.2010.09.008>
- Tsekouras K, Chatzistamoulou N, Kounetas K, Broadstock DC (2016) Spillovers, path dependence and the productive performance of European transportation sectors in the presence of technology heterogeneity. *Technol Forecast Soc Chang* 102:261–274
- Tsekouras K, Chatzistamoulou N, Kounetas K (2017) Productive performance, technology heterogeneity and hierarchies: who to compare with whom. *Int J Prod Econ* 193:465–478
- Wang Q, Su B, Zhou P, Chiu CR (2016) Measuring total-factor CO<sub>2</sub> emission performance and technology gaps using a non-radial directional distance function: a modified approach. *Energy Econ* 56:475–482
- Wang X, Zhang M, Nathwani J, Yang F (2019) Measuring environmental efficiency through the lens of technology heterogeneity: a comparative study between China and the G20. *Sustainability* 11(2):461. <https://doi.org/10.3390/su11020461>
- Wei Y, Li Y, Wu M, Li Y (2019) The decomposition of total-factor CO<sub>2</sub> emission efficiency of 97 contracting countries in Paris agreement. *Energy Econ* 78:365–378. <https://doi.org/10.1016/j.eneco.2018.11.028>
- Xian Y, Yang K, Wang K, Wei Y, Huang Z (2019) Cost-environment efficiency analysis of construction industry in China: a materials balance approach. *J Clean Prod* 221:457–468
- York R, Rosa EA, Dietz T (2003) STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. *Ecol Econ* 46(3):351–365
- Zaim O, Taskin F (2000) Environmental efficiency in carbon dioxide emissions in the OECD: a non-parametric approach. *J Environ Manage* 58(2):95–107
- Zofio JL, Prieto AM (2001) Environmental efficiency and regulatory standards: the case of CO<sub>2</sub> emissions from OECD industries. *Resour Energy Econ* 23(1):63–83
- Zhang N, Kong F, Choi Y (2014) Measuring sustainability performance for China: a sequential generalized directional distance function approach. *Econ Model* 41:392–397
- Zhang Y, Li H, Li Y, Zhou LA (2010) FDI spillovers in an emerging market: the role of foreign firms' country origin diversity and domestic firms' absorptive capacity. *Strateg Manag J* 31(9):969–989
- Zhou P, Ang BW, Wang H (2012) Energy and CO<sub>2</sub> emission performance in electricity generation: a non-radial directional distance function approach. *Eur J Oper Res* 221(3):625–635

# Data Mining Approach in Personnel Selection: The Case of the IT Department



Ezgi Demir, Rana Elif Dinçer, Batuhan Atasoy, and Sait Erdal Dinçer

**Abstract** Data mining studies have been frequently included in the literature recently. Data mining can be applied in every field, especially in banking, marketing, customer relationship management, investment and portfolio management. In the literature, the problem of personnel selection has been examined with the help of multi-criteria decision-making techniques. In this study, it has been aimed to apply data mining techniques in the field of human resources where relatively little has been used. The features of a large-scale construction company have been determined according to the competencies specified in the information technologies department announcement. The candidates were ranked according to these attributes. While ranking, accuracy values have been compared by using basic algorithms of data mining techniques. While applying the process steps, the necessary data pre-processing techniques have been applied to candidates who entered incomplete or incorrect information during the application process. Basically, the decision trees algorithm gave the highest accuracy. Also, random forest, adaboost, gradient boosting, and xgboost algorithms have been tried. In addition, it has found the attributes that should be looked at first in the application features. The high number of data enabled machine learning to learn information more easily and to weigh the existing criteria easily. With this study, it has been aimed to obtain a more objective result by weighting with machine learning algorithms instead of weighting the personnel selection problem with multi-criteria decision-making methodology.

---

E. Demir (✉)

Department of Management Information Systems, Piri Reis University, Tuzla, Turkey

R. E. Dinçer

Department of Business Informatics, Marmara University, Göztepe, Turkey

B. Atasoy

Department of Mechanical Engineering, Piri Reis University, Tuzla, Turkey

e-mail: [batasoy@pirireis.edu.tr](mailto:batasoy@pirireis.edu.tr)

S. E. Dinçer

Department of Econometrics, Marmara University, Göztepe, Turkey

e-mail: [edincer@marmara.edu.tr](mailto:edincer@marmara.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_22](https://doi.org/10.1007/978-3-030-85254-2_22)

In addition, it is an extremely difficult process to interview candidates for recruitment under the current Covid-19 pandemic conditions that the whole world and our country are struggling with. Online conversations take a lot of time. With this study, it has been aimed to provide optimization by automating the process by weighting the features related to the existing data in the process. The study has been done in the WEKA and Python program.

**Keywords** Personnel Selection · Data Mining · Information Technologies Department · Artificial Intelligence · Boosting Algorithms

## 1 Introduction

In today's information society, analyzing data in a meaningful way is a vital issue for businesses. It is an important development to be able to record every transaction made, to create important and meaningful data, and to analyze this data. The large amount of data obtained with the help of information and communication technologies provides convenience to companies, countries, institutions, and organizations in various fields. It has also created various new requirements. In order to adapt to this developing new generation business, understanding data mining methods have been used to discover hidden information in the data. Data mining methods now have been actively used in almost every department of businesses. One of these departments is the human resources, which deals with the manpower and is the main resource of the company. Data mining can be used in the human resources department in areas such as performance management, personnel planning, analysis of various risks, analysis and reporting of important performance criteria. It is also an important tool that can be used to modernize and digitize business processes. In the human resources process, it is extremely important to evaluate the personnel selection according to the qualifications of the personnel and to select the right personnel. A large number of applications have reached the departments. Increasing unemployment processes and decreasing employment rate create increases in applications for positions. Numerous transactions are carried out regarding personnel in human resources departments. Undoubtedly, one of the longest processes is the selection of the personnel to be placed in a position and the negotiation process. In fact, these interviews can continue in 3–4 stages and multiple interviews can be made. All of these processes keep the human resources department busy and prevent them from concentrating on other jobs. This situation reduces the efficiency of the department. At the same time, the process of recruiting personnel requires coordinated work with the managers of the departments where the personnel will be recruited. In this covid-19 pandemic, etc., processes, recruitment and management of new employees in crisis situations have become even more difficult. Staff interviews have been conducted online. Until today, solutions have been found for personnel selection practices in human resources with multi-criteria decision-making methodologies. However, multi-criteria decision-making methodology has judgments that vary from

person to person. With this study, data mining algorithms were used by bringing a new perspective to personnel selection. With machine learning techniques, one of the data mining techniques, the feature values of the applicants were extracted and the personnel selected according to these criteria were made. With this study conducted in the Covid-19 process, a more objective study has been tried to be put forward, away from the interview with candidates. With this study, a model proposal has been presented for the information technology department, which meets the demands of the enterprise in objective conditions in personnel selection. Personnel selection for the information technology department requires a large number of qualities. Multi-criteria decision-making methodology has been used for situations where the number of criteria are high. However, when the number of criteria are high, it becomes difficult to decompose decision criteria that take values from 0–1. In this study, unlike other studies, personnel selection was made with a data mining methodology, which provides more objective results.

This study was conducted within the scope of the application process for the announcement of a holding established in Istanbul in 1987 with code BİM-SRM. In this announcement, it has been planning to hire an information processing officer for the mining department. Seventy-two people applied for the announcement. The criteria required in the application listed have been as follows. Information with 26 criteria was requested. These are application date, announcement code, position, department\_name, name, surname, scope of the candidate, candidate's hometown, candidate's age, candidate's gender, candidate's nationality, candidate's military status information, whether the candidate has a travel impediment, the candidate's employment barrier, the candidate's marital status, the name of the university the candidate graduated from, the university department the candidate graduated from, the candidate's starting year, the candidate's license completion year, the number of foreign languages the candidate knows and which ones (English, German, Russian results have been obtained), MsOffice Information on the programs, the candidate's programming language knowledge (results such as Java, C++ were obtained.), the candidate's database information (answers such as Mysql, Oracle, Rdbms), the candidate's knowledge of SAP Modules, and finally, the candidate's work experience on a yearly basis were obtained.

Candidates have been classified according to their criterion degrees according to data mining methodologies and suitable candidates have been determined. Thus, each candidate was not interviewed separately. During the Covid-19 pandemic process, it has been determined that it is the right method to carry out the process with data mining methodologies for recruitment process. Instead of determining the criteria to be considered in the selection of candidates by human resources employees, data mining methodology has determined the weights of the criteria with learning algorithms and classified candidates with high accuracy. As a result of the study, if deemed necessary, the company has the opportunity to choose by making online interviews with the suitable candidates.

## 2 Modern Approaches in Human Resources Management

Throughout history, the concepts of human and resource management come to the fore wherever there is human and organized labor. From this point of view, it can be understood that this concept is not only an understanding developed as a result of management theories, but a concept that emerged in ancient times and as a result of natural processes (Griffin, 2006). However, as in every field, some updates have been needed in the field of human resources management in order to talk about an effective operation in line with the twenty-first century. There are factors such as the rapid progress in the field of information technologies in the last 20 years, thus the economic globalization has reached extraordinary dimensions and the competitive environment has become global as a result of the limits eliminated by this globalization. Therefore, it has today made it compulsory for both individual employees and organizations more generally to modernize themselves. The concept of owning, holding, developing, and using intellectual capital, which has become a critical factor of superiority in such a global competitive environment, is largely linked to the effectiveness of human resources management. For this reason, modern human resources management approaches have become indispensable for organizations today.

## 3 Data Mining Application Areas in Organizations

In the growing information society, almost all obstacles to information exchange have been removed. Global knowledge has revealed another dimension of global competition. This developing global environment not only offers a renewed competition environment in terms of workforce quality and quantity. At the same time, data have been gathered from a global environment and evolved into a form that can be compiled. This has led to an increase in the importance given to data and the race to obtain meaningful and useful information from complex data. In such a large and complex information pool, people have come to need forward-looking systems that are difficult to predict as well as internal and external information in order to gain a competitive advantage. Data warehouse, information management, and data mining, which are the three main areas that have developed in response to these needs, are generally aimed at getting more information from data (Silahdaroğlu, 2013). Data mining technology supports companies in finding useful information between large-scale data and mining information. The greatest convenience that data mining technology offers companies is that it helps identify similar trends and patterns of behavior across data sets. This feature is widely used in marketing activities, especially for target markets. Another great help is that it makes relationships that were not visible at first easily visible. For example, a company can analyze the products it sells, design its future campaigns accordingly, discover the links between the products it sells, and develop a marketing tactic based on those links. In today's dynamic business world, there is a high risk that people's decisions will be wrong

or that decision-makers' information will become obsolete. And of course, there is no space for errors in an area where such fragile and momentary changes occur. The only way to reduce such risks is to use decision support solutions that provide knowledge-based solutions. The meaningful data obtained through data mining techniques helps organizations to make strategic decisions correctly, to better manage their risks, and to be innovative. Due to such driving forces, data mining technology has become one of the most important requirements for organizations today.

## 4 Methodology and Application

In this study, a data mining study was carried out on the CVs of 72 candidates sent to the human resources department for the IT staff to be employed in the mining department. In this context, 5 candidates out of 72 candidates provided incomplete information and were removed from the system during data processing. The remaining 67 candidates were interviewed with human resources personnel, and the candidates' compliance with the starting date requested by the company and their communication skills were evaluated, taking into account their compliance with the company policy and company ownership. Being able to perform data mining applications in human resources does not fall within the scope of an application that can only be done on data. In addition to the qualifications of the individuals, the pre-interview data are an important result in classifying the candidates' data.

In this study, the results of candidates' starting work as a result of the interview with the human resources officer were coded as "suitable" and "not suitable". Candidates' eligibility status has been classified with data mining techniques. Figure 1 shows the data mining application steps. In this context, the needs of the enterprise were analyzed first. As a result of the online interviews, it has been observed that many interviews are held in the human resources process in enterprises, these processes both create a long waiting period for candidates who want to work, depending on the intensity of candidate application, and create an intense loss of work and workforce in the recruitment process for human resources personnel. In the human resources process, the candidates' suitability for the job is evaluated first, and face-to-face interviews are held in the last step, and the suitability of the candidates is evaluated in accordance with the starting date of the company. Thanks to the model, all candidates have been determined as "suitable" and "not suitable", thanks to the determination of their starting and communication skills rather than wasting time while qualitatively separating the candidates, and the candidates and their attributes (attributes) are taught to the machine with the machine learning algorithms used in data mining. "suitable" or "not suitable" will be separated.

Adhering to the data mining application steps, after the needs of the enterprise were understood, the data of the enterprise was taken in the excel file during the applications, and the candidates with the missing data in the data were eliminated in order to prepare for the system. In this context, incompatibilities of the data against the English character have been corrected in order to be loaded into the system. In the

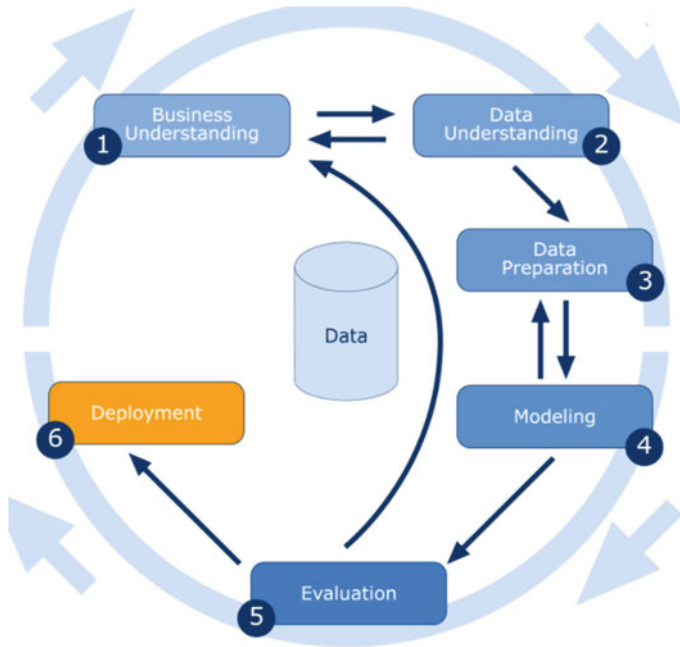


Fig. 1 Data mining implementation steps

data on the military status of the candidates, exemption, postponed, made, and absent (for female candidates) are disaggregated depending on gender. While the number of foreign languages known by the candidates shows a value of 0,1,2, the value 0,1 in the MSOffice programs information is indicated as the candidate’s knowledge of the programs: 1, and the status of not knowing: 0. Likewise, the candidate’s programming knowledge (Java, C++), database knowledge, and sap modular knowledge are coded as “1”, while the candidate’s ignorance is coded as “0”. Another data editing process is numerical quantities such as 0,3,10,13,14 in the experience period of the candidates. Candidates’ work experience increases as the numerical quantity increases in the year column (expressed as the attribute value), their knowledge and experience increase. After the relevant data preparation was defined, when the model was set up during the modeling phase, when the 26 different attribute values of the candidates were examined, it was observed that the application dates, application code, position, department, country, gender information did not affect the results of the candidates’, “suitable” or “not suitable”. Another factor that makes a difference in the model is that it is seen that the marital status of the candidates does not affect the result, but only the travel disability attribute, so there is no need to use the marital status variable in the model. Because there is a travel disability variable overlapping with the marital status variable. Another factor that makes a difference is that the year of starting the candidates’ undergraduate or related departments does not have an effect on the result as an attribute, but instead, the year of graduation of the candidates





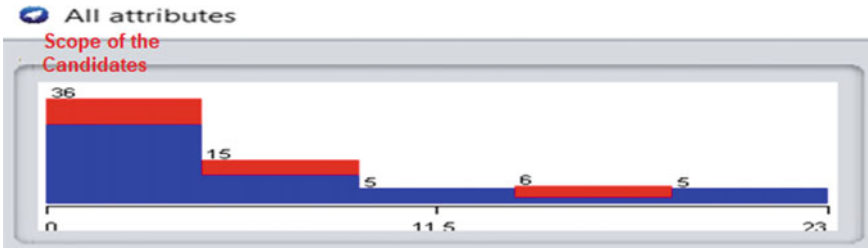


Fig. 4 Distribution of the candidates’ fields

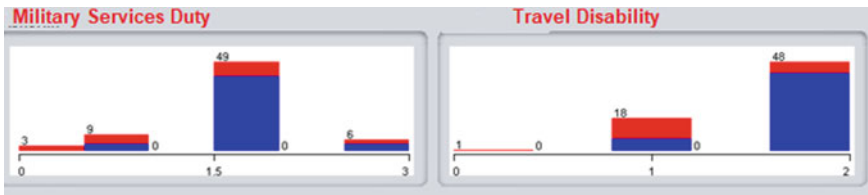


Fig. 5 Distribution of “military service” and “travel disability” status of the candidates

In Fig. 5, the suitability of the candidates according to their military service status and travel disability has been examined. In the case of military service, it is seen that the most suitable ones from four categories (done, postponed, exempted-not, and not) have done their military service, then exempted-no, postponed, and in the last case those who did not.

Figure 5 also looks at the travel disability status of the candidates, and it is seen that the blue density of the candidates who do not have travel disability is more appropriate in the current two situations (there is a disability, there is no disability). However, it cannot be said that every candidate without travel disability is eligible, as there are red “unsuitable” candidates in the histogram on the far right without travel barriers. This shows that the data is distributed homogeneously.

In Fig. 6, the candidates’ suitability according to the name-departments of the university has been examined.

As can be seen in Fig. 6, the eligibility of the candidates according to the university names and departments is homogeneously distributed. Blue indicates conformity, red indicates unfit. Due to the high number of applications from computer, hardware,

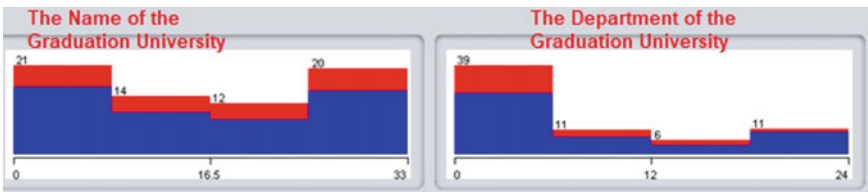
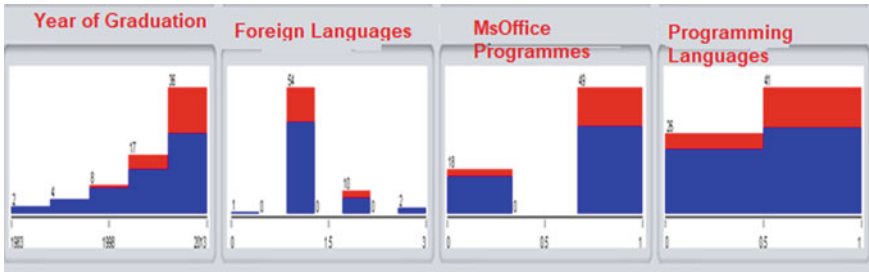


Fig. 6 University names and departments of candidates



**Fig. 7** Licensing expiry year, number of foreign languages, office program, and programming information of candidates

and information departments, the number of eligible candidates on the far right side of the University Department chart is as high, while there are many non-eligible candidates from the same department due to the high competition for this application announcement.

In Fig. 7, license expiry year, number of foreign languages, office program, and programming information of the candidates are taken into consideration. As the graduation years of the candidates approached today, their eligibility increased, and the unsuitability status due to their recent graduates and their inexperience increased. There are four categories in the number of foreign languages which are known. There is a suitable candidate who does not speak (0: no spoken, 1: speak one language, 2: speak 2 languages, 3: speak 3 languages). While the majority of suitable candidates who know one language are in the majority, most of those who are not suitable are those who speak one language.

In Fig. 7, there are two categories as those who know office programs and those who do not. The most suitable ones are seen in the category of those who know. Likewise, those who know java and c++ have also made a significant difference in compliance. In Fig. 8, the candidates' database information, sap modular knowledge, work experience, and candidate eligibility in the last case have been examined.

According to what is given in Fig. 8, the suitability of those who know the database has increased by far, the suitability of those with sap modular knowledge increases by far. The eligibility of those with more experience in work experience has increased.



**Fig. 8** Candidates' database information, sap modular knowledge, work experience, and candidate eligibility in the last case

The eligibility of newly graduated candidates is as high. Data are homogeneously distributed. In the last case, 49 of the candidates are eligible, while 18 of them are not.

Prediction of candidates was made by applying data mining. It was modeled in the python program.

The decision tree algorithm has been selected and 80% of the data have been parsed as learning data, while 20% of the data have been separated as test data.

The results of running the data have been shown in Fig. 9. Relation part shows the value where the file name is registered. Instances shows the number of rows (number of people) to be analyzed. The attribute values show the feature values of the lines. Here, candidates have 20 properties. In machine learning, learning algorithms have been divided into either 67% training, 33% test data or 80% training, and 20% test data. A higher learning rate always gives better results. Therefore, 80% learning is reserved as 20% test data. Train 80% expression is the learning rate. Remain (the rest) is test data. And also the model has been tried for random forest algorithm. The model results have been shown in Fig. 10. Another model has been tried for Adaboost algorithm in Fig. 11. Gradient boosting and Xgboost algorithms have been tried in Figs. 12 and 13, respectively.

The accuracy of all the models has been shown in Fig. 14. Accuracy percentages have been given according to the cross validation evaluations of the models. Best results have been seen when cross validation is 4 and 8 in Fig. 14.

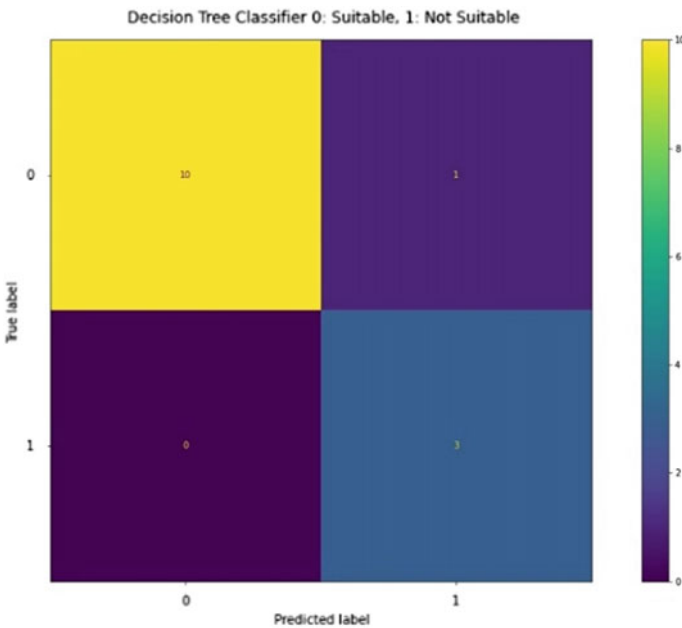


Fig. 9 Results of the data for decision tree algorithm

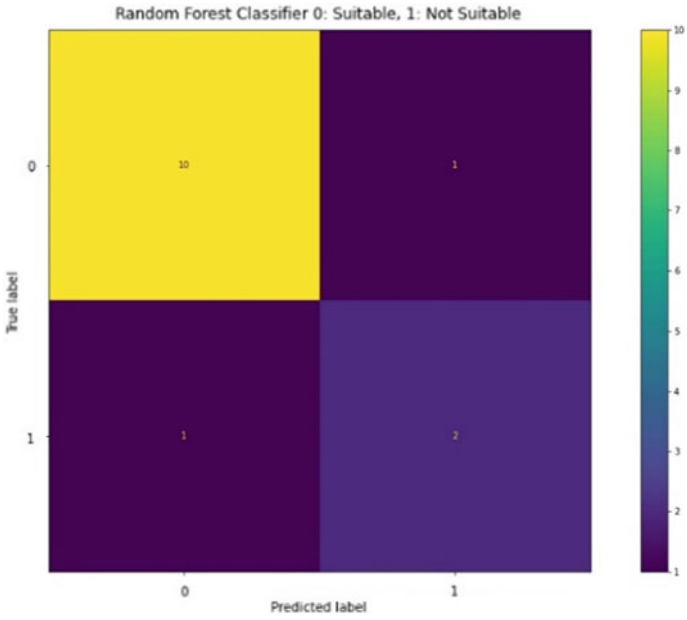


Fig. 10 Results of the model for random forest algorithm

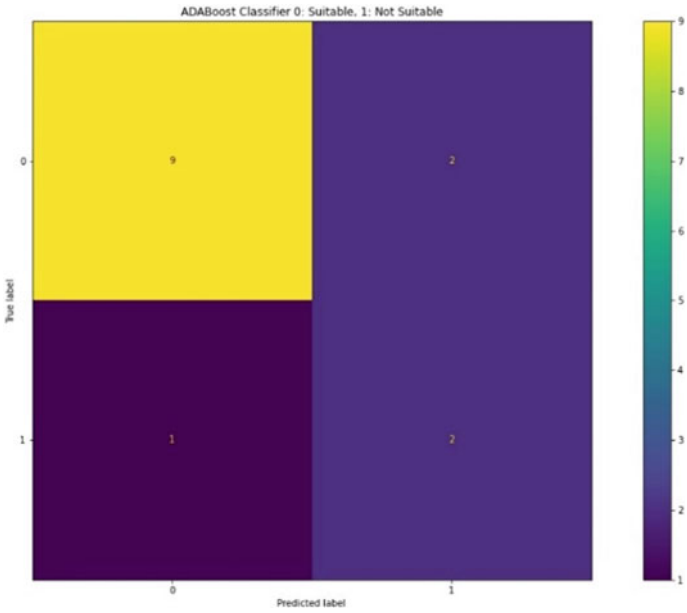


Fig. 11 Results of the model for ADABOOST algorithm

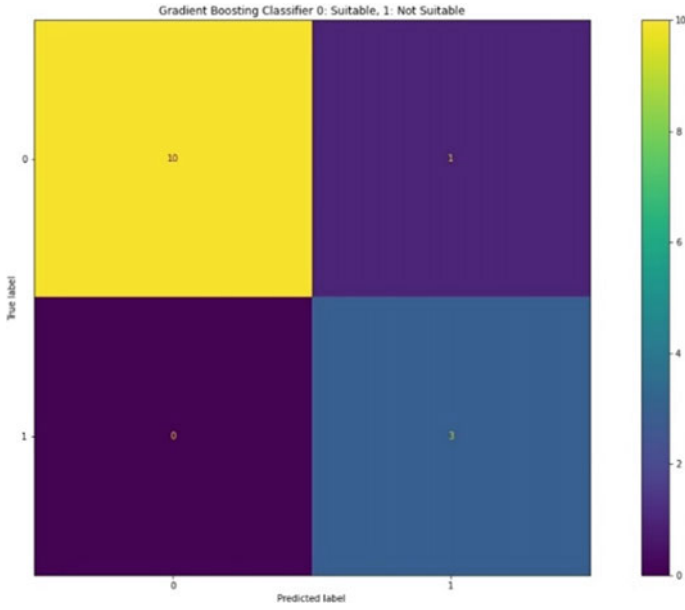


Fig. 12 Results of the model for Gradient Boosting algorithm

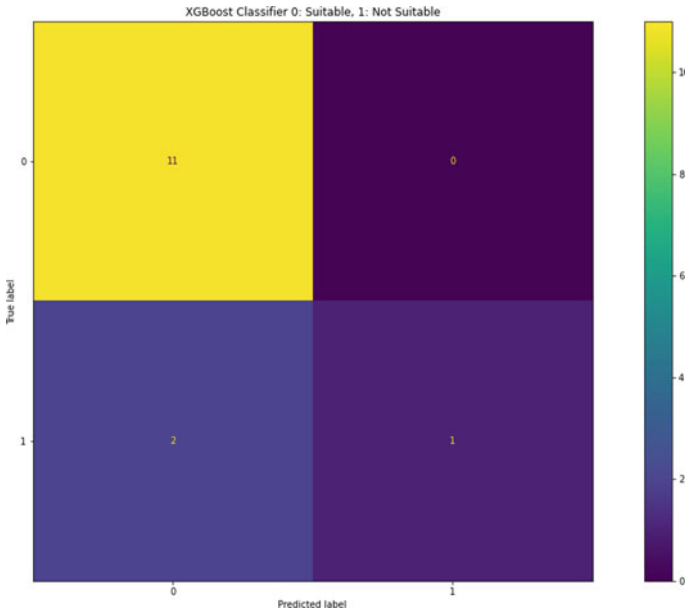


Fig. 13 Results of the model for XGBoost algorithm

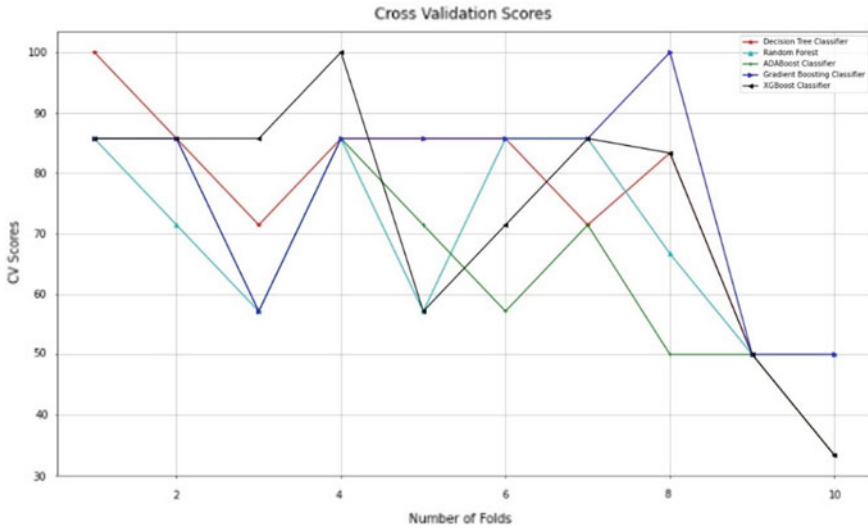


Fig. 14 Cross validation scores

## 5 Conclusion

Data mining is an important step in the process of finding information, which is known as Information and Data Discovery in the literature, which reveals data with various methods (Şeker, 2016). Data Mining is a research and iterative process that separates and filters the patterns in the knowledge discovery process and prepares them for the next step (Coşlu, 2013).

When examined in detail, it can be seen that the data mining process includes the following steps (Akpınar, 2000):

1. Creating a data set or obtaining it from a ready source,
2. Organizing the data (minimizing repetition, filling missing parts in the data, etc.),
3. Selecting the appropriate data mining technique and algorithm for the project,
4. Application and interpretation of the model,
5. Reporting and presenting the obtained information.

The first thing that matters in a data mining application is to set a roadmap based on the goals and framework of the project. In this study, the recruitment processes of a construction company during the pandemic process were examined. During the pandemic process, it is extremely difficult to have long and multiple interviews with candidates. For this reason, employee selection was made with data mining techniques. First, the characteristics of the candidates for recruitment were determined. Later, the candidates were examined individually according to their eligibility. Finally, by applying the machine learning methodology, a hierarchical structure was

created according to the criteria that should be considered in the recruitment process of the candidates. Candidates who were eligible for recruitment were classified.

With this study, it is aimed to gain a different methodology to the subject of personnel selection than the multi-criteria decision-making methodology, which is frequently used in the literature. Expert opinions and quantitative value judgments play an important role in multi-criteria decision-making methodology. In data mining and machine learning techniques, more objective results can be obtained as the criteria are determined by the machine. In this respect, the study is aimed to shed light on multi-criteria and data science studies in the literature.

## References

- Akpınar H (2000) Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İşletme Fakültesi Dergisi, 1–22
- Coşlu E (2013) Veri Madenciliği. Akdeniz Üniversitesi, Antalya: XV. Akademik Bilişim Konferansı Bildiriler
- Griffin R (2006) Principles of management. Houghton Mifflin Company
- Silahdaroğlu G (2013) Veri Madenciliği . İstanbul: Papatya Yayınları
- Şeker E (2016) Veri Madenciliği Araçları YBS Ansiklopedi Cilt 3, Sayı 4. YBS Ansiklopedi Cilt 3, Sayı 4



# A Possibilistic Programming Approach to Portfolio Optimization Problem Under Fuzzy Data



Pejman Peykani, Mohammad Namakshenas, Mojtaba Nouri, Neda Kavand, and Mohsen Rostamy-Malkhalifeh

**Abstract** Investment portfolio optimization problem is an important issue and challenge in the investment field. The goal of portfolio optimization problem is to create an efficient portfolio that incurs the minimum risk to the investor across different return levels. It should be noted that in many real cases, financial data are tainted by uncertainty and ambiguity. Accordingly, in this study, the fuzzy portfolio optimization model using possibilistic programming is presented that is capable to be used in the presence of fuzzy data and linguistic variables. Three objectives including the return, the systematic risk, and the non-systematic risk are considered to propose the fuzzy portfolio optimization model. Finally, the possibilistic portfolio optimization model is implemented in a real case study from the Tehran stock exchange to show the efficacy and applicability of the proposed approach.

**Keywords** Portfolio optimization problem · Fuzzy optimization · Stock return · Systematic risk · Non-Systematic risk · Possibilistic programming

## 1 Introduction

Portfolio optimization (PO) problem has been a practically important challenge and issue in financial markets ((Markowitz 1952); (Lobo et al. 2007); (Kalayci et al. 2019); (Ahmadi-Javid and Fallah-Tafti 2019)). PO problem is concerned with choosing an optimal and efficient portfolio strategy that can strike a trade-off between maximizing investment return and minimizing investment risk ((Konno

---

P. Peykani · M. Namakshenas · M. Nouri  
School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran  
e-mail: [m\\_namakshenas@ind.iust.ac.ir](mailto:m_namakshenas@ind.iust.ac.ir)

M. Nouri  
e-mail: [mojtaba\\_nouri@ind.iust.ac.ir](mailto:mojtaba_nouri@ind.iust.ac.ir)

N. Kavand · M. Rostamy-Malkhalifeh (✉)  
Department of Mathematics, Faculty of Basic Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran

and Suzuki 1995); (Cesarone et al. 2013); (Björk et al. 2014); (Guo et al. 2019); (Salah et al. 2020)). The important point that should be considered for modeling investment PO problem is the uncertainty of financial data ((Ghahtarani and Najafi 2013, 2018); (Huang 2017); (Peykani and Mohammadi 2018); (Li et al. 2019); (Peykani et al. 2020)). It should be noted that for proposing uncertain portfolio optimization (UPO) models, according to the nature and type of uncertainty, the popular uncertain programming approaches such as stochastic optimization (SO) ((Land et al. 1993); (Sueyoshi 2000); (Cooper et al. 2002); (Wu et al. 2013); (Zha et al. 2016)), fuzzy optimization (FO) ((Zadeh 1978); (Azadeh and Kokabi 2016); (Peykani et al. 2018a, 2019a, 2019b, 2021); (Peykani and Mohammadi 2018); (Seyed Esmaeili et al. 2019); (Peykani and Gheidar-Kheljani 2020)), and robust optimization (RO) ((Soyster 1973); (Ben-Tal and Nemirovski 2000); (Bertsimas and Sim 2004); (Ghassemi et al. 2017); (Namakshenas et al. 2017); (Ghassemi 2019); (Namakshenas and Pishvaei 2019); (Peykani and Roghanian 2015); (Peykani and Mohammadi 2018); (Peykani et al. 2018, 2019, 2019d; Peykani et al. 2020)) can be utilized.

The goal of this paper is to present the possibilistic portfolio optimization (PPO) model that is capable to be implemented under fuzzy data. It should be explained that a possibilistic programming approach is employed for dealing with the uncertainty and ambiguity of financial data. The possibilistic programming is an applicable and powerful approach for dealing with the epistemic uncertainty that is caused by the absence or lack of knowledge about the exact value of model parameters in fuzzy mathematical programming (FMP) (Naderi et al., 2016). Notably, to demonstrate the applicability of the proposed fuzzy portfolio optimization model, the PPO approach is implemented in a Tehran stock exchange (TSE).

The rest of this paper is organized as follows. The modeling of the deterministic portfolio optimization model will be explained in Sect. 2. Then, the possibilistic portfolio optimization model using possibilistic programming will be proposed in Sect. 3. The proposed PPO model is employed for the sample test from the Tehran stock exchange in Sect. 4. Finally, conclusions, as well as some future research directions, will be introduced in Sect. 5.

## 2 Deterministic Portfolio Optimization Model

In this section, the deterministic portfolio optimization (DPO) model will be introduced. It should be noted that three aspects of stock including the rate of return, the systematic risk, and the non-systematic risk are considered in the DPO model. The nomenclatures of the paper are introduced as follows:

$j$	the indices of stocks $j = 1, \dots, n$
$t$	the indices of periods $t = 1, \dots, T$
$R_E$	the expected return of portfolio

(continued)

(continued)

- $j$  the indices of stocks  $j = 1, \dots, n$
- $R_j$  the average return of  $j$ th stock
- $R_{tj}$  the return of  $j$ th stock in  $t$ th period
- $R_M$  the return of the market
- $\sigma_M^2$  the variance of the market
- $\beta_E$  the expected beta of the portfolio
- $\beta_j$  the beta of  $j$ th stock
- $A_j$  the minimum level of total fund which can be invested in the  $j$ th stock
- $B_j$  the maximum level of total fund which can be invested in the  $j$ th stock
- $\varpi_j$  the weight of  $j$ th stock in the portfolio
- $\Omega_t$  the value of the non-systematic risk of the portfolio in  $t$ th period
- $\tau_j$  a binary variable which will be one if  $j$ th stock is selected and zero otherwise

Investment risk can be decomposed into two components including systematic risk and non-systematic risk. Systematic risk includes that part of the risk which depends on market variability and is unavoidable.

Beta ( $\beta$ ) sensitivity coefficient is one of the most popular systematic risk measures. The beta coefficient describes the sensitivity of the share return to the market portfolio return. In other words, beta is a measure of the volatility of share with the overall market and is obtained from Eq. (1) as follows:

$$\beta = \frac{Cov(R_j, R_M)}{\sigma_M^2} \tag{1}$$

Non-systematic risk indicates that a part of the investment risk can be eliminated by diversification. The absolute deviation (AD) is a non-systematic risk measure introduced by Konno & Yamazaki ((Konno and Yamazaki 1991)) for the first time. The definition of absolute deviation is as given in Eq. (2):

$$AD = |R_j - R_E| = \begin{cases} R_j - R_E \text{ if } R_j > R_E; \\ R_E - R_j \text{ if } R_j \leq R_E. \end{cases} \tag{2}$$

Now, the mean-absolute deviation-beta (MADB) model for portfolio optimization problem is presented as Model (3):

$$\begin{aligned} & \text{Min} \frac{1}{T} \sum_{t=1}^T \Omega_t \tag{3} \\ & \text{S.t.} \sum_{j=1}^n R_j \varpi_j \geq R_E \end{aligned}$$

$$R_E - \sum_{j=1}^n R_{tj} \varpi_j \leq \Omega_t, \forall t$$

$$\sum_{j=1}^n R_{tj} \varpi_j - R_E \leq \Omega_t, \forall t$$

$$\sum_{j=1}^n \beta_j \varpi_j \leq \beta_E$$

$$\sum_{j=1}^n \varpi_j = 1$$

$$A_j \tau_j \leq \varpi_j \leq B_j \tau_j, \forall j$$

$$\tau_j \in \{0, 1\}, \forall j$$

$$\Omega_t \geq 0, \forall t$$

$$\varpi_j \geq 0, \forall j$$

It should be noted that in Model (3),  $A_j$  and  $B_j$  are the lower and the upper bounds, respectively, for each stock and  $\varpi_j$  is the portion of the investment portfolio that is assigned to each stock. Accordingly, the limit of investment for each stock as a common financial market constraint is considered in the proposed portfolio optimization model.

### 3 Possibilistic Portfolio Optimization Model

In this section, the possibilistic portfolio optimization (PPO) model under fuzzy data will be presented. It should be explained that the return and the beta have a trapezoidal fuzzy distribution  $\tilde{R}(R^1, R^2, R^3, R^4)$  and  $\tilde{\beta}(\beta^1, \beta^2, \beta^3, \beta^4)$  with the condition of  $R^1 < R^2 < R^3 < R^4$  and  $\beta^1 < \beta^2 < \beta^3 < \beta^4$ . Now, a possibilistic programming approach and chance-constrained programming (CCP) will be applied to deal with fuzzy data in the MADB model as follows:

$$\text{Min } \frac{1}{T} \sum_{t=1}^T \Omega_t \quad (4)$$

$$\begin{aligned}
 & \text{S.t. Pos} \left\{ \sum_{j=1}^n \tilde{R}_j \varpi_j \geq R_E \right\} \geq \delta \\
 & \text{Pos} \left\{ R_E - \sum_{j=1}^n \tilde{R}_{tj} \varpi_j \leq \Omega_t \right\} \geq \delta, \forall t \\
 & \text{Pos} \left\{ \sum_{j=1}^n \tilde{R}_{tj} \varpi_j - R_E \leq \Omega_t \right\} \geq \delta, \forall t \\
 & \text{Pos} \left\{ \sum_{j=1}^n \tilde{\beta}_j \varpi_j \leq \beta_E \right\} \geq \delta \\
 & \sum_{j=1}^n \varpi_j = 1 \\
 & A_j \tau_j \leq \varpi_j \leq B_j \tau_j, \forall j \\
 & \tau_j \in \{0, 1\}, \forall j \\
 & \Omega_t \geq 0, \forall t \\
 & \varpi_j \geq 0, \forall j
 \end{aligned}$$

Then, by applying the possibility measure and chance-constrained programming, converting fuzzy chance-constraints into their equivalent crisp ones in one special confidence level ( $\delta$ ) is done as follows:

$$\begin{aligned}
 & \text{Min} \frac{1}{T} \sum_{t=1}^T \Omega_t \tag{5} \\
 & \text{S.t.} \sum_{j=1}^n \left( (\delta)R_j^3 + (1 - \delta)R_j^4 \right) \varpi_j \geq R_E \\
 & R_E - \sum_{j=1}^n \left( (\delta)R_{tj}^3 + (1 - \delta)R_{tj}^4 \right) \varpi_j \leq \Omega_t, \forall t \\
 & \sum_{j=1}^n \left( (1 - \delta)R_{tj}^1 + (\delta)R_{jt}^2 \right) \varpi_j - R_E \leq \Omega_t, \forall t \\
 & \sum_{j=1}^n \left( (1 - \delta)\beta_j^1 + (\delta)\beta_j^2 \right) \varpi_j \leq \beta_E
 \end{aligned}$$

(continued)

(continued)

$$\text{Min} \frac{1}{T} \sum_{t=1}^T \Omega_t \tag{5}$$

$$\sum_{j=1}^n \varpi_j = 1$$

$$A_j \tau_j \leq \varpi_j \leq B_j \tau_j, \forall j$$

$$\tau_j \in [0, 1], \forall j$$

$$\Omega_t \geq 0, \forall t$$

$$\varpi_j \geq 0, \forall j$$

Finally, the possibilistic mean-absolute deviation-beta (PMADB) model is proposed as Model (5) that can be employed by investors for portfolio optimization under fuzzy data and linguistic variables.

### 4 Experimental Results

In this section, the possibilistic portfolio optimization model will be implemented for a real-world case study from the Tehran stock exchange. Accordingly, the data set for five stocks are extracted from TSE. Tables 1, 2, 3, 4, and 5 show the data set for beta and return of five stocks under trapezoidal fuzzy number:

**Table 1** Fuzzy data set for beta

Stocks	$\beta^1$	$\beta^2$	$\beta^3$	$\beta^4$
Stock 1	0.1704	0.20235	0.22365	0.2556
Stock 2	0.6392	0.75905	0.83895	0.9588
Stock 3	0.5544	0.65835	0.72765	0.8316
Stock 4	1.0216	1.21315	1.34085	1.5324
Stock 5	0.0968	0.11495	0.12705	0.1452

**Table 2** Fuzzy data set for monthly return—first period

Stocks	$R^1$	$R^2$	$R^3$	$R^4$
Stock 1	18.13	21.53	23.79	27.19
Stock 2	5.91	7.01	7.75	8.86
Stock 3	4.06	4.82	5.33	6.09
Stock 4	9.43	11.20	12.38	14.15
Stock 5	8.93	10.60	11.72	13.39

**Table 3** Fuzzy Data Set for Monthly Return—Second Period

Stocks	$R^1$	$R^2$	$R^3$	$R^4$
Stock 1	-5.10	-6.06	-6.70	-7.65
Stock 2	-2.77	-3.29	-3.63	-4.15
Stock 3	8.80	10.45	11.55	13.20
Stock 4	3.21	3.81	4.21	4.81
Stock 5	23.69	28.13	31.09	35.53

**Table 4** Fuzzy data set for monthly return—third period

Stocks	$R^1$	$R^2$	$R^3$	$R^4$
Stock 1	10.40	12.35	13.65	15.60
Stock 2	14.37	17.07	18.86	21.56
Stock 3	24.59	29.20	32.27	36.88
Stock 4	24.45	29.03	32.09	36.68
Stock 5	14.05	16.68	18.44	21.07

**Table 5** Fuzzy data set for monthly return—average

Stocks	$R^1$	$R^2$	$R^3$	$R^4$
Stock 1	7.81	9.27	10.25	11.71
Stock 2	5.84	6.93	7.66	8.76
Stock 3	12.48	14.82	16.38	18.72
Stock 4	12.36	14.68	16.23	18.55
Stock 5	15.56	18.47	20.42	23.33

Now, after collecting data, the possibilistic mean-absolute deviation-beta model will be run. The results of the PMADB model that is presented in Model (5) for five confidence levels including 0, 25, 50, 75, and 100% are introduced in Table 6:

As can be seen in Table 6, by increasing the confidence level from 0 to 100%, the objective functions including mean, absolute deviation, and beta get worse. Also, illustrative results show that the proposed PMADB model is effective for portfolio optimization in the presence of fuzzy data.

## 5 Conclusions

In this study, an uncertain portfolio optimization model is presented that is capable to be used under fuzzy environment. It should be explained that in the proposed portfolio optimization model, three objectives including mean, absolute deviation, and

**Table 6** The results of the PMADB model

PMADB Model		Confidence Levels				
		0%	25%	50%	75%	100%
Weight	Stock 1	0.30	0.30	0.30	0.30	0.30
	Stock 2	0.10	0.10	0.10	0.00	0.00
	Stock 3	0.30	0.30	0.30	0.30	0.30
	Stock 4	0.00	0.00	0.00	0.10	0.10
	Stock 5	0.30	0.30	0.30	0.30	0.30
Portfolio	Return	17.004	16.473	15.943	16.299	15.738
	AD	1.413	1.79	2.181	2.664	3.226
	Beta	0.739	0.774	0.808	0.819	0.852

beta as well as investment constraint are considered. Also, the possibilistic programming and chance-constrained programming approaches are employed to deal with uncertainty. For future studies, data envelopment analysis approach (Seyed Esmaeili, 2014; Peykani et al., 2018c; Peykani and Mohammadi, 2018, 2019, 2020; Seyed Esmaeili and Rostamy-Malkhalifeh, 2018), machine learning models (Park et al., 2014; Ban et al., 2018; Shahhosseini et al., 2019, 2020, 2019; Paiva et al., 2019), and game theory ( Migdalas, 2002; Sadeghi and Zandieh, 2011; Esmaeili et al., 2015) can be applied for presenting investment portfolio optimization approach.

## References

- Ahmadi-Javid A, Fallah-Tafti M (2019) Portfolio optimization with entropic value-at-risk. *Eur J Oper Res* 279(1):225–241
- Azadeh A, Kokabi R (2016) Z-number DEA: a new possibilistic DEA in the context of Z-numbers. *Adv Eng Inform* 30(3):604–617
- Ban GY, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. *Manage Sci* 64(3):1136–1154
- Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Math Program* 88(3):411–424
- Bertsimas D, Sim M (2004) The price of robustness. *Oper Res* 52(1):35–53
- Björk T, Murgoci A, Zhou XY (2014) Mean–variance portfolio optimization with state-dependent risk aversion. *Math Financ: Int J Math Stat Financ Econ* 24(1):1–24
- Cesarone F, Scozzari A, Tardella F (2013) A new method for mean-variance portfolio optimization with cardinality constraints. *Ann Oper Res* 205(1):213–234
- Cooper WW, Deng H, Huang Z, Li SX (2002) Chance constrained programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis. *J Oper Res Soc* 53(12):1347–1356
- Esmaeili M, Bahrini A, Shayanrad S (2015) Using game theory approach to interpret stable policies for Iran's oil and gas common resources conflicts with Iraq and Qatar. *J Ind Eng Int* 11(4):543–554
- Ghahtarani A, Najafi AA (2013) Robust goal programming for multi-objective portfolio selection problem. *Econ Model* 33:588–592



- Ghahtarani A, Najafi AA (2018) Robust optimization in portfolio selection by m-MAD model approach. *Econom Comput Econom Cybernet Stud Res* 52(1):279–291
- Ghassemi A, Hu M, Zhou Z (2017) Robust planning decision model for an integrated water system. *J Water Resour Plan Manag* 143(5):05017002
- Ghassemi A (2019) System of systems approach to develop an energy-water nexus model under uncertainty. Doctoral Dissertation, University of Illinois
- Guo X, Chan RH, Wong WK, Zhu L (2019) Mean–variance, mean–VaR, and mean–CVaR models for portfolio selection with background risk. *Risk Manage* 21(2):73–98
- Huang X (2017) A review of uncertain portfolio selection. *J Intell Fuzzy Syst* 32(6):4453–4465
- Kalayci CB, Ertenlice O, Akbay MA (2019) A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. *Expert Syst Appl* 125:345–368
- Konno H, Suzuki KI (1995) A mean-variance-skewness portfolio optimization model. *J Oper Res Soc Jpn* 38(2):173–187
- Konno H, Yamazaki H (1991) Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Manage Sci* 37(5):519–531
- Land KC, Lovell CK, Thore S (1993) Chance-constrained data envelopment analysis. *Manag Decis Econ* 14(6):541–554
- Li B, Sun Y, Aw G, Teo KL (2019) Uncertain portfolio optimization problem under a minimax risk measure. *Appl Math Model* 76:274–281
- Lobo MS, Fazel M, Boyd S (2007) Portfolio optimization with linear and fixed transaction costs. *Ann Oper Res* 152(1):341–365
- Markowitz H (1952) Portfolio selection. *J Financ* 7(1):77–91
- Migdalas A (2002) Applications of game theory in finance and managerial accounting. *Oper Res Int J* 2(2):209–241
- Naderi MJ, Pishvae MS, Torabi SA (2016) Applications of fuzzy mathematical programming approaches in supply chain planning problems. In: *Fuzzy logic in its 50th year*. Springer, Cham, pp 369–402
- Namakshenas M, Pishvae MS, Mahdavi Mazdeh M (2017) Event-driven and attribute-driven robustness. *Iran J Oper Res* 8(1):78–90
- Namakshenas M, Pishvae MS (2019) Data-driven robust optimization. Robust and constrained optimization: methods and applications. Nova Science Publishers, Inc, pp 1–40
- Paiva FD, Cardoso RTN, Hanaoka GP, Duarte WM (2019) Decision-making for financial trading: a fusion approach of machine learning and portfolio selection. *Expert Syst Appl* 115:635–655
- Park J, Lim J, Lee W, Ji S, Sung K, Park K (2014) Modern probabilistic machine learning and control methods for portfolio optimization. *Int J Fuzzy Log Intell Syst* 14(2):73–83
- Peykani P, Gheidar-Kheljani J (2020) Performance appraisal of research and development projects value-chain for complex products and systems: the fuzzy three-stage DEA approach. *J New Res Math* 6(25):41–58
- Peykani P, Mohammadi E (2019) Performance measurement of decision making units with network structure in the presence of undesirable output. *J New Res Math* 5(17):157–166
- Peykani P, Mohammadi E (2020) Window network data envelopment analysis: an application to investment companies. *Int J Ind Math* 12(1):89–99
- Peykani P, Roghanian E (2015) The application of data envelopment analysis and robust optimization in portfolio selection problem. *J Oper Res Its Appl* 12(44):61–78
- Peykani P, Mohammadi E, Pishvae MS, Rostamy-Malkhalifeh M, Jabbarzadeh A (2018a) A novel fuzzy data envelopment analysis based on robust possibilistic programming: possibility, necessity and credibility-based approaches. *RAIRO-Oper Res* 52(4–5):1445–1463
- Peykani P, Mohammadi E, Seyed Esmaeili FS (2018c) Measuring performance, estimating most productive scale size, and benchmarking of hospitals using DEA approach: a case study in Iran. *Int J Hosp Res* 7(2):21–41
- Peykani P, Mohammadi E, Rostamy-Malkhalifeh M, Hosseinzadeh Lotfi F (2019a) Fuzzy data envelopment analysis approach for ranking of stocks with an application to Tehran stock exchange. *Adv Math Financ Appl* 4(1):31–43

- Peykani P, Mohammadi E, Emrouznejad A, Pishvaei MS, Rostamy-Malkhalifeh M (2019b) Fuzzy data envelopment analysis: an adjustable approach. *Expert Syst Appl* 136:439–452
- Peykani P, Mohammadi E, Seyed Esmaeili FS (2019d) Stock evaluation under mixed uncertainties using robust DEA model. *J Qual Eng Prod Optim* 4(1):73–84
- Peykani P, Mohammadi E (2018) Robust data envelopment analysis with hybrid uncertainty approaches and its applications in stock performance measurement. In: *Proceedings of The 14th International Conference on Industrial Engineering*. Iran
- Peykani P, Mohammadi E (2018) Fuzzy network data envelopment analysis: a possibility approach. In: *Proceedings of The 3th International Conference on Intelligent Decision Science*. Iran
- Peykani P, Mohammadi E (2018) Interval network data envelopment analysis model for classification of investment companies in the presence of uncertain data. *J Ind Syst Eng* 11(Special issue: 14th International Industrial Engineering Conference):63–72
- Peykani P, Mohammadi E (2018) Portfolio selection problem under uncertainty: a robust optimization approach. In: *Proceedings of The 3th International Conference on Intelligent Decision Science*. Iran
- Peykani P, Mohammadi E, Sadjadi SJ, Rostamy-Malkhalifeh M (2018) A robust variant of radial measure for performance assessment of stock. In: *Proceedings of The 3th International Conference on Intelligent Decision Science*. Iran
- Peykani P, Seyed Esmaeili FS, Hosseinzadeh Lotfi F, Rostamy-Malkhalifeh M (2019) Estimating most productive scale size in DEA under uncertainty. In: *Proceedings of The 11th National Conference on Data Envelopment Analysis*. Iran
- Peykani P, Mohammadi E, Jabbarzadeh A, Rostamy-Malkhalifeh M, Pishvaei MS: A novel two-phase robust portfolio selection and optimization approach under uncertainty: A case study of Tehran stock exchange. *Plos One* 15(10), e0239810 (2020)
- Peykani P, Mohammadi E, Farzipoor Saen R, Sadjadi SJ, Rostamy-Malkhalifeh M (2020) Data envelopment analysis and robust optimization: a review. *Expert Syst* 37(4):e12534
- Peykani P, Mohammadi E, Emrouznejad A (2021) An adjustable fuzzy chance-constrained network DEA approach with application to ranking investment firms. *Expert Syst Appl* 166:113938
- Sadeghi A, Zandieh M (2011) A game theory-based model for product portfolio management in a competitive market. *Expert Syst Appl* 38(7):7919–7923
- Salah HB, Gannoun A, Ribatet M (2020) Conditional Mean-Variance and Mean-Semivariance models in portfolio optimization. *J Stat Manag Syst* 1–24
- Seyed Esmaeili FS (2014) The efficiency of MSBM model with imprecise data (interval). *Int J Data Envel Anal* 2(1):343–350
- Seyed Esmaeili FS, Rostamy-Malkhalifeh M (2018) Using interval data envelopment analysis (IDEA) to performance assessment of hotel in the presence of imprecise data. In: *Proceedings of The 3th International Conference on Intelligent Decision Science*. Iran
- Seyed Esmaeili FS, Rostamy-Malkhalifeh M, Hosseinzadeh Lotfi F (2019) The possibilistic Malmquist productivity index with fuzzy data. In: *Proceedings of The 11th National Conference on data envelopment analysis*. Iran
- Shahhosseini M, Hu G, Archontoulis SV (2020) Forecasting corn yield with machine learning ensembles. *Front Plant Sci* 11:1120
- Shahhosseini M, Martinez-Feria RA, Hu G, Archontoulis SV (2019) Maize yield and nitrate loss prediction with machine learning algorithms. *Environ Res Lett* 14(12):124026
- Shahhosseini M, Hu G, Pham H (2019) Optimizing ensemble weights for machine learning models: a case study for housing price prediction. In: *Proceedings of INFORMS International Conference on Service Science*. Springer, Cham, pp 87–97
- Soyster AL (1973) Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper Res* 21(5):1154–1157
- Sueyoshi T (2000) Stochastic DEA for restructure strategy: an application to a Japanese petroleum company. *Omega* 28(4):385–398

- Wu C, Li Y, Liu Q, Wang K (2013) A stochastic DEA model considering undesirable outputs with weak disposability. *Math Comput Model* 58(5–6):980–989
- Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1(1):3–28
- Zha Y, Zhao L, Bian Y (2016) Measuring regional efficiency of energy and carbon dioxide emissions in China: a chance constrained DEA approach. *Comput Oper Res* 66:351–361

# A Hybrid Fuzzy MCDM Approach for ESCO Selection



Nilsen Kundakcı

**Abstract** In recent years, energy efficiency has become an important issue due to the growing energy demands of countries and firms. High costs of energy supplies and environmental issues are the main problems around the world and lead to a global effort for saving energy. Besides, there is a growing interest in providing energy services to achieve energy and environmental goals. Therefore, new companies called Energy Service Companies (ESCOs) providing energy services to energy users started to operate in the world market. In this context, it is important for the firms to choose the right Energy Service Company that will enable them to save energy and to assist in their energy efficiency projects. In this paper, a hybrid fuzzy Multi-Criteria Decision-Making (MCDM) approach is proposed for Energy Service Company selection process of a textile firm. This approach is based on fuzzy Stepwise Weight Assessment Ratio Analysis (SWARA) and fuzzy Measurement Alternatives and Ranking according to the Compromise Solution (MARCOS) methods. Firstly, the weights of decision criteria are determined by using fuzzy SWARA method. Later, Energy Service Company alternatives are evaluated with the help of fuzzy MARCOS method, and the best alternative for the textile firm is determined.

**Keywords** Fuzzy MCDM · Fuzzy SWARA · Fuzzy MARCOS · ESCO selection

## 1 Introduction

Improving energy efficiency is an essential component of a sustainable energy policy and has an important role in developing climate change mitigation strategies. Energy Service Companies (ESCOs) aid firms to improve energy performance, ensure savings, reduce energy costs, and finance or help to arrange financing for the operation of an energy system. In addition to these, they deliver energy efficiency improvement measures, and assist the firms in developing a roadmap to execute the

---

N. Kundakcı (✉)

Department of Business Administration, Pamukkale University, Denizli, Turkey

e-mail: [nilsenk@pau.edu.tr](mailto:nilsenk@pau.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

389

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_24](https://doi.org/10.1007/978-3-030-85254-2_24)

energy efficiency projects. Energy efficiency projects involve the issues such as identifying opportunities to actively reduce energy usage in a facility, energy saving equipment that yields improvement in energy efficiency, and saving in both economic and environmental terms. Firms prefer to work with ESCOs to save energy, money, and to be more environmentally conscious. They play a crucial role in improving energy efficiency goals of the firms as they have the necessary experience to provide solutions achieving significant energy cost reductions. In this paper, a hybrid fuzzy MCDM (Multi-Criteria Decision-Making) approach is proposed for ESCO selection problem of a textile firm operating in Denizli, Turkey. In this approach, fuzzy SWARA (Step-wise Weight Assessment Ratio Analysis) method is used to calculate the weights of decision criteria, and ESCO alternatives are evaluated with the help of fuzzy MARCOS (Measurement Alternatives and Ranking according to the Compromise Solution) method. By this way, the ranking of the ESCO alternatives is obtained, and the most appropriate alternative is selected for the textile firm.

The reason why fuzzy SWARA and fuzzy MARCOS methods are preferred in this paper is concerning with the reason that conventional MCDM methods are inadequate to deal with ambiguity and uncertainty in the decision-making process. In classical SWARA and MARCOS methods, crisp values are used while determining the criteria weights and evaluating the alternatives. It is assumed that they are known precisely. However, crisp values are insufficient to model real-life decision-making problems. For this reason, in this paper, fuzzy hybrid MCDM approach is proposed in which the ratings of alternatives and weights of criteria are evaluated by linguistic variables represented by triangular fuzzy numbers to overcome the deficiency in the conventional MCDM methods.

In this paper, fuzzy SWARA method is proposed in determination of the criteria weights as it is one of the new methods being used to evaluate criteria. In the literature, different methods such as fuzzy AHP (Analytic Hierarchy Process) and fuzzy MACBETH (Measuring Attractiveness through a Categorical-Based Evaluation Technique) methods were also proposed to determine the criteria weights. The reason for choosing fuzzy SWARA method in the study is that it includes simpler calculations compared to fuzzy AHP and fuzzy MACBETH methods and the result can be reached in a shorter time. Additionally, fuzzy SWARA method is a newer method than the other methods. On the other hand, fuzzy SWARA method can facilitate group decisions, and more than one expert opinion has been considered in determining the criteria weights in this paper.

Fuzzy MARCOS method that uses the ratio method, and the reference point method has some advantages over other fuzzy MCDM methods. For instance, it considers fuzzy reference points through the fuzzy ideal and fuzzy anti-ideal solution at the beginning of the construction of initial matrix. Moreover, fuzzy MARCOS proposes a novel approach for determining utility functions and its aggregation. At the same time, it ensures that many criteria and alternatives are considered by maintaining the stability of the method.

The main contribution of this paper to the literature is the integrated use of fuzzy SWARA and fuzzy MARCOS methods, which are two relatively new fuzzy multi-criteria decision-making methods. Additionally, the proposed hybrid approach aims

to take advantage of the strengths of the fuzzy SWARA and fuzzy MARCOS methods, and it will assist decision-makers of the firms in evaluating the alternatives, and in selecting the best one that satisfy the needs and expectations of their firms. On the other hand, this paper discusses the fuzzy MCDM problem of a textile firm with the proposed approach and provides guidance for researchers interested in real-life applications of fuzzy MCDM methods.

The organization of this paper is as follows. In the second section, firstly, fuzzy sets and then fuzzy numbers are summarized and algebraic operations with fuzzy numbers are also demonstrated. In the third and fourth sections, fuzzy SWARA and fuzzy MARCOS methods are explained, respectively, and the steps of these methods are summarized. Literature reviews of these methods are also mentioned. In the fifth section, application of the proposed method in a textile firm for ESCO selection is presented. In the last section, results of the proposed hybrid approach are discussed, and suggestions for future research are offered.

## 2 Fuzzy Sets and Fuzzy Numbers

The fuzzy set theory first introduced in 1965 by Zadeh for dealing with imprecision of human thought. Zadeh defines fuzzy set as a class of objects whose grades of membership are continuous. Fuzzy set is defined with a membership function which assigns to each object a grade of membership ranging from 0 to 1 (Zadeh, 1965).

A fuzzy number  $\tilde{A}$  is a convex normalized fuzzy set. It is an extension of a real number and does not refer to a single value, but to an associated set of possible weight values between 0 and 1. Although various types of fuzzy numbers are defined, usually triangular and trapezoidal fuzzy numbers are preferred in practice for ease of calculation. The membership function of a triangular fuzzy number  $\tilde{A} = (l, m, u)$  can be seen in Eq. 1.

$$\mu_{\tilde{A}} = \begin{cases} 0, & x < l, \\ (x - l)/(m - l), & l \leq x \leq m, \\ (u - x)/(u - m), & m \leq x \leq u, \\ 0, & x > u \end{cases} \tag{1}$$

$\tilde{A} = (l_1, m_1, u_1)$  and  $\tilde{B} = (l_2, m_2, u_2)$  are two positive triangular fuzzy numbers, and  $k$  is a positive real number. Then, basic operations of fuzzy numbers  $\tilde{A}$  and  $\tilde{B}$  can be summarized as in Eqs. 2–7 (Kaufmann and Gupta, 1988).

$$\tilde{A} + \tilde{B} = (l_1 + l_2, m_1 + m_2, u_1 + u_2) \tag{2}$$

$$\tilde{A} - \tilde{B} = (l_1 - u_2, m_1 - m_2, u_1 - l_2) \tag{3}$$

$$\tilde{A}xk = (l_1.k, m_1.k, u_1.k) \quad (4)$$

$$\tilde{A}x\tilde{B} = (l_1.l_2, m_1.m_2, u_1.u_2) \quad (5)$$

$$\tilde{A}/\tilde{B} = (l_1/u_2, m_1/m_2, u_1/l_2) \quad (6)$$

$$\tilde{A}^{-1} = (l_1, m_1, u_1)^{-1} = (1/u_1, 1/m_1, 1/l_1) \quad (7)$$

### 3 Fuzzy SWARA Method

The Stepwise Weight Assessment Ratio Analysis (SWARA) method was firstly proposed by Kersulienė et al. (2010) as a tool to determine the criteria weights in MCDM problems. Later, it was extended to fuzzy SWARA by Mavi et al. in 2017. After it was extended to fuzzy SWARA, it was used for determining the criteria weights in the literature by various authors. For instance, Zarbakhshnia et al. (2018) used fuzzy SWARA and fuzzy COPRAS methods to evaluate and select third-party reverse logistics provider. Additionally, Perçin (2019) proposed an integrated fuzzy SWARA and fuzzy Axiomatic Design approach to select outsourcing provider. Petrović et al. (2019) evaluated suppliers with three fuzzy MCDM methods. Fuzzy SWARA has been used to determine the criteria weights, and supplier alternatives have been evaluated with fuzzy ARAS (Additive Ratio Assessment), fuzzy TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), and fuzzy WASPAS (Weighted Aggregated Sum Product Assessment). In this scope, Kaya and Erginel (2020) integrated hesitant fuzzy SWARA and hesitant fuzzy sustainable Quality Function Deployment methods for sustainable airport quality and design. Lastly, Agarwal et al. (2020) proposed a hybrid fuzzy SWARA and fuzzy WASPAS approach and evaluated solutions to overcome humanitarian supply chain management barriers.

The fuzzy SWARA method is performed by following the steps below (Mavi et al., 2017; Perçin, 2019):

**Step 1:** Sort the criteria in descending order depending on their importance degrees. For instance, the most important criterion is ranked as first, and the least important one is assigned as last.

**Step 2:** Each decision-maker in the decision committee expresses the relative importance of criterion  $j$  in relation to the previous one ( $j-1$ ), beginning from the second criterion. This ratio  $s_j$  is called the comparative importance of average value (Kersulienė et al., 2010). The fuzzy comparison scale demonstrated in Table 1 can be used while evaluating the criteria.

Then aggregated values of decision-makers' judgments for evaluation criteria can be obtained with the help of Eq. 8

**Table 1** The fuzzy comparison scale for the evaluation of criteria (Yazdani, 2011)

Linguistic variable	Response scale
Very Low (VL)	(0, 0, 0.25)
Low (L)	(0, 0.25, 0.50)
Medium (M)	(0.25, 0.50, 0.75)
High (H)	(0.50, 0.75, 1)
Very High (VH)	(0.75, 1, 1)

$$\tilde{s}_j = (s_j^l, s_j^m, s_j^u) = \left( \frac{\sum_{k=1}^K s_{jk}^l}{K}, \frac{\sum_{k=1}^K s_{jk}^m}{K}, \frac{\sum_{k=1}^K s_{jk}^u}{K} \right) \tag{8}$$

Here  $\tilde{s}_j$  indicates the aggregated judgment of decision-makers ( $k = 1, 2, 3 \dots K$ ) for the criterion  $j$ .

**Step 3.** Obtain coefficient  $\tilde{k}_j$  values with the help of Eq. 9.

$$\tilde{k}_j = \begin{cases} \tilde{1}, & j = 1 \\ \tilde{s}_j + 1, & j > 1 \end{cases} \tag{9}$$

**Step 4.** Compute fuzzy weights of criteria  $\tilde{q}_j$  by using Eq. 10.

$$\tilde{q}_j = \begin{cases} \tilde{1}, & j = 1 \\ \frac{\tilde{q}_{j-1}}{\tilde{k}_j}, & j > 1 \end{cases} \tag{10}$$

**Step 5.** Calculate the final weights of criteria  $\tilde{w}_j$  via Eq. 11.

$$\tilde{w}_j = \frac{\tilde{q}_j}{\sum_{k=1}^n \tilde{q}_k} \tag{11}$$

here  $\tilde{w}_j = (\tilde{w}_j^l, \tilde{w}_j^m, \tilde{w}_j^u)$ .

### 4 Fuzzy MARCOS Method

MARCOS method was firstly introduced by Stevic et al. in 2020. It is based on a utility function which represents the position of the alternative regarding the ideal and anti-ideal solution. This method is based on a compromise ranking in relation to the ideal and anti-ideal solutions, and the best alternative is the one closest to the ideal and the furthest from the anti-ideal reference point (Stevic et al., 2020).



In this regard, Stankovic et al. (2020) extended the MARCOS method to fuzzy environment and develop fuzzy MARCOS method. Although fuzzy MARCOS method is a new method, it has been applied to different fields in the literature. It can be used in many studies in the future due to its ease of calculations and effective results. Besides, Ilieva et al. (2020) proposed to use fuzzy MARCOS method for cloud service selection. Mijajlović et al. (2020) integrated fuzzy MARCOS and fuzzy FUCOM (Full Consistency Method) to determine the competitiveness of spa-centers for achieving sustainability.

Steps of the fuzzy MARCOS method can be given as follows (Stankovic et al., 2020):

**Step 1:** Construct the initial fuzzy matrix  $\tilde{A} = [\tilde{x}_{ij}]_{m \times n}$  as seen in Eq. (12)

$$\tilde{A} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1n} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \tilde{x}_{m1} & \tilde{x}_{m2} & \cdots & \tilde{x}_{mn} \end{bmatrix} \tag{12}$$

**Step 2:** Construct the extended initial fuzzy matrix ( $\tilde{X}$ ). This matrix is obtained by adding fuzzy ideal  $\tilde{A}(ID)$  and fuzzy anti-ideal  $\tilde{A}(AI)$  solutions to the initial fuzzy matrix.  $\tilde{A}(AI)$  values are added to the first row of the initial fuzzy matrix, while  $\tilde{A}(ID)$  values are added to the last row.

The fuzzy  $\tilde{A}(ID)$  is the best alternative while the fuzzy  $\tilde{A}(AI)$  is the worst alternative. Based on type of the criteria,  $\tilde{A}(ID)$  and  $\tilde{A}(AI)$  are defined by Eqs. 13 and 14.

$$\tilde{A}(ID) = \max_i \tilde{x}_{ij} \text{ if } j \in B \text{ and } \min_i \tilde{x}_{ij} \text{ if } j \in C \tag{13}$$

$$\tilde{A}(AI) = \min_i \tilde{x}_{ij} \text{ if } j \in B \text{ and } \max_i \tilde{x}_{ij} \text{ if } j \in C \tag{14}$$

In Eqs. 13 and 14,  $B$  and  $C$  indicate benefit and cost criteria, respectively.

**Step 3:** Normalize the extended initial fuzzy matrix and obtain normalized fuzzy matrix  $\tilde{N} = [\tilde{n}_{ij}]_{m \times n}$  with the help of Eqs. 15 and 16.

$$\tilde{n}_{ij} = (n_{ij}^l, n_{ij}^m, n_{ij}^u) = \left( \frac{x_{ij}^l}{x_{id}^u}, \frac{x_{ij}^m}{x_{id}^u}, \frac{x_{ij}^u}{x_{id}^u} \right) \text{ if } j \in B \tag{15}$$

$$\tilde{n}_{ij} = (n_{ij}^l, n_{ij}^m, n_{ij}^u) = \left( \frac{x_{id}^l}{x_{ij}^u}, \frac{x_{id}^l}{x_{ij}^m}, \frac{x_{id}^l}{x_{ij}^l} \right) \text{ if } j \in C \tag{16}$$

Here  $x_{ij}^l, x_{ij}^m, x_{ij}^u$  and  $x_{id}^l, x_{id}^m, x_{id}^u$  are the elements of the matrix  $\tilde{X}$ .

**Step 4:** Construct the weighted normalized fuzzy matrix  $\tilde{V} = [\tilde{v}_{ij}]_{m \times n}$  with the help of Eq. 17.

$$\tilde{v}_{ij} = (v_{ij}^l, v_{ij}^m, v_{ij}^u) = \tilde{n}_{ij} \otimes \tilde{w}_j = (n_{ij}^l \cdot w_j^l, n_{ij}^m \cdot w_j^m, n_{ij}^u \cdot w_j^u) \tag{17}$$

**Step 5:** Calculate fuzzy  $\tilde{S}_i$  matrix by using Eq. 18.

$$\tilde{S}_i = \sum_{j=1}^n \tilde{v}_{ij} \tag{18}$$

Here  $\tilde{S}_i (s_i^l, s_i^m, s_i^u)$  indicates the sum of elements of the weighted normalized fuzzy matrix  $\tilde{V}$ .

**Step 6:** Calculate the utility degrees of alternatives  $\tilde{K}_i$  by using Eqs. 19 and 20.

$$\tilde{K}_i^- = \frac{\tilde{S}_i}{\tilde{S}_{ai}} = \left( \frac{s_i^l}{s_{ai}^u}, \frac{s_i^m}{s_{ai}^m}, \frac{s_i^u}{s_{ai}^l} \right) \tag{19}$$

$$\tilde{K}_i^+ = \frac{\tilde{S}_i}{\tilde{S}_{id}} = \left( \frac{s_i^l}{s_{id}^u}, \frac{s_i^m}{s_{id}^m}, \frac{s_i^u}{s_{id}^l} \right) \tag{20}$$

**Step 7:** Calculate fuzzy matrix  $\tilde{T}_i$  with the help of Eq. 21.

$$\tilde{T}_i = \tilde{t}_i = (t_i^l, t_i^m, t_i^u) = \tilde{K}_i^- + \tilde{K}_i^+ = (k_i^{-l} + k_i^{+l}, k_i^{-m} + k_i^{+m}, k_i^{-u} + k_i^{+u}) \tag{21}$$

Later a new fuzzy number  $\tilde{D}$  is obtained by using Eq. 22.

$$\tilde{D} = (d^l, d^m, d^u) = \max_i \tilde{t}_{ij} \tag{22}$$

Then fuzzy number  $\tilde{D}$  is defuzzified by using Eq. 23 and  $df_{crisp}$  number is obtained.

$$df_{crisp} = \frac{l + 4m + u}{6} \tag{23}$$

**Step 8:** Determine utility functions regarding to the ideal  $f(\tilde{K}_i^+)$  and anti-ideal  $f(\tilde{K}_i^-)$ .

$$f(\tilde{K}_i^+) = \frac{\tilde{K}_i^-}{df_{crisp}} = \left( \frac{k_i^{-l}}{df_{crisp}}, \frac{k_i^{-m}}{df_{crisp}}, \frac{k_i^{-u}}{df_{crisp}} \right) \tag{24}$$

$$f(\tilde{K}_i^-) = \frac{\tilde{K}_i^+}{df_{crisp}} = \left( \frac{k_i^{+l}}{df_{crisp}}, \frac{k_i^{+m}}{df_{crisp}}, \frac{k_i^{+u}}{df_{crisp}} \right) \tag{25}$$

Later,  $\tilde{K}_i^-, \tilde{K}_i^+, f(\tilde{K}_i^+), f(\tilde{K}_i^-)$  values are defuzzified.

**Step 9:** Determine the utility functions of alternatives  $f(K_i)$  by using Eq. 26.

$$f(K_i) = \frac{K_i^+ + K_i^-}{1 + \frac{1-f(K_i^+)}{f(K_i^+)} + \frac{1-f(K_i^-)}{f(K_i^-)}} \quad (26)$$

**Step 10:** Rank the alternatives depending on the final values of their utility functions. The alternative with the highest value of utility function will be the best one.

## 5 Application

In this application part, Energy Service Company (ESCO) selection problem of a textile firm operating in Denizli, Turkey is considered. For solving this problem, a hybrid fuzzy MCDM is proposed. Steps of the proposed hybrid approach are summarized on Fig. 1.

Firstly, a committee of decision-makers is formed in the textile firm. This committee consists of three decision-makers: energy manager  $DM_1$ , mechanical engineer  $DM_2$ , and general manager  $DM_3$ . Later, they define the alternatives and evaluation criteria. After a preliminary research, decision committee determines three ESCO alternatives as  $A_1$ ,  $A_2$ , and  $A_3$ . The decision committee defines six criteria as follows:

- **C<sub>1</sub> Knowledge:** Knowledge of the ESCO about strategies, projects, and markets.
- **C<sub>2</sub> References:** Reference firms that ESCO works with in the market.
- **C<sub>3</sub> Fee:** The payment made to the ESCO due to its consultancy.
- **C<sub>4</sub> Experience:** Experience of ESCO in the related field.
- **C<sub>5</sub> Certificates:** The number and content of certificates owned by ESCO.
- **C<sub>6</sub> Communication skills:** Communication ability of ESCO in the organization level.

After the decision committee determines the alternatives and criteria, the weights of the criteria are determined with the help of Fuzzy SWARA method. In the first step of fuzzy SWARA, decision-makers rank the criteria in descending order based on their importance degrees. Later, they evaluate the criteria by using the linguistic variables in Table 1. The evaluation results of three decision-makers are presented in Table 2.

Then, aggregated values of decision-makers' judgments for the evaluation criteria are obtained via Eq. 8. These aggregated values called as comparative importance of average value  $\tilde{s}_j$  are obtained in the second step and demonstrated in Table 3.

In the third step of SWARA method,  $\tilde{k}_j$  values are obtained with the help of Eq. 9. Then in the fourth step, fuzzy weights of criteria  $\tilde{q}_j$  are calculated by using Eq. 10. In the last step, final weights of criteria  $\tilde{w}_j$  are calculated via Eq. 11. These values are shown in Table 4, respectively.

The final weights  $\tilde{w}_j$  of the six criteria are demonstrated in the last row of Table 4. According to these values, the most important criterion is Knowledge ( $C_1$ ) and

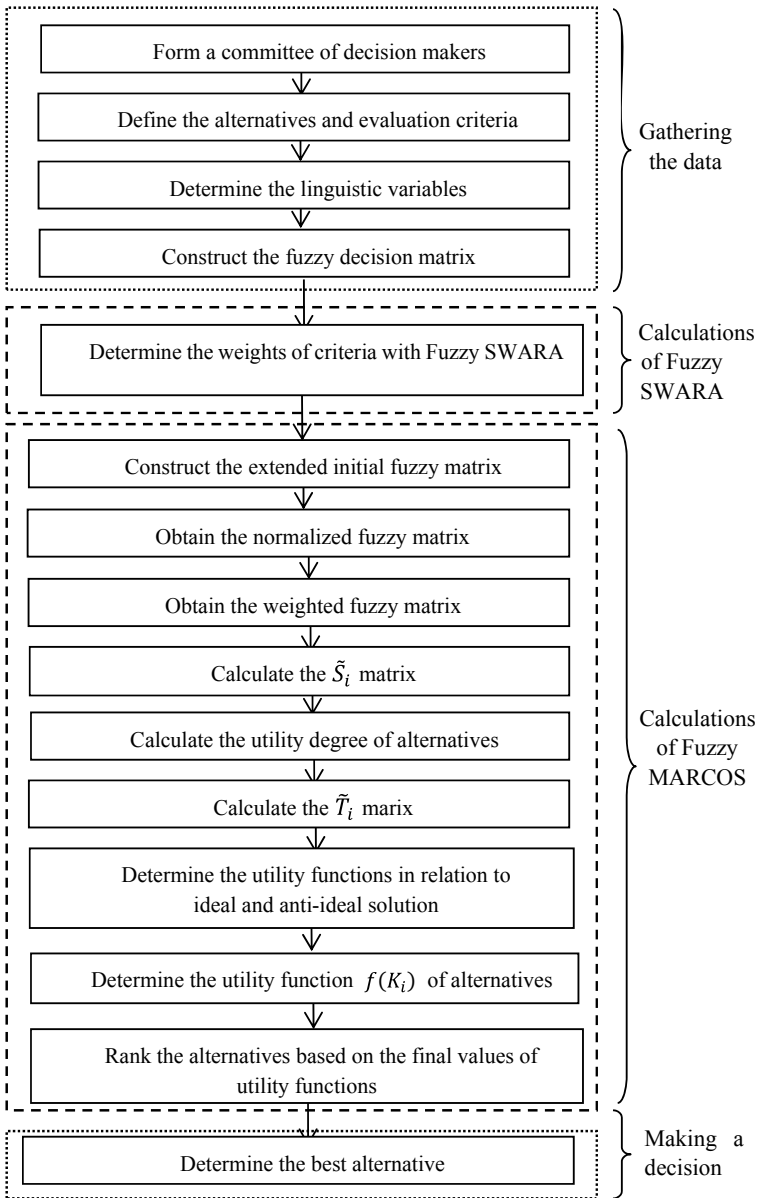


Fig. 1 Steps of the hybrid fuzzy MCDM approach

**Table 2** Evaluation results for the criteria

Rating with linguistic variables				Rating with TFNs		
	DM <sub>1</sub>	DM <sub>2</sub>	DM <sub>3</sub>	DM <sub>1</sub>	DM <sub>2</sub>	DM <sub>3</sub>
C <sub>1</sub>						
C <sub>2</sub>	M	VL	L	(0.25, 0.50, 0.75)	(0, 0, 0.25)	(0, 0.25, 0.50)
C <sub>3</sub>	M	L	M	(0.25, 0.50, 0.75)	(0, 0.25, 0.50)	(0.25, 0.50, 0.75)
C <sub>4</sub>	H	M	M	(0.50, 0.75, 1)	(0.25, 0.50, 0.75)	(0.25, 0.50, 0.75)
C <sub>5</sub>	M	VH	H	(0.25, 0.50, 0.75)	(0.75, 1, 1)	(0.50, 0.75, 1)
C <sub>6</sub>	H	M	M	(0.75, 1, 1)	(0.25, 0.50, 0.75)	(0.25, 0.50, 0.75)

**Table 3** The comparative importance of average value  $\tilde{s}_j$

	Criteria		$\tilde{s}_j$	
C <sub>1</sub>	Knowledge			
C <sub>2</sub>	References	0.083	0.250	0.500
C <sub>3</sub>	Fee	0.167	0.417	0.667
C <sub>4</sub>	Experience	0.333	0.583	0.833
C <sub>5</sub>	Certificates	0.500	0.750	0.917
C <sub>6</sub>	Communication skills	0.417	0.667	0.833

**Table 4**  $\tilde{k}_j$ ,  $\tilde{q}_j$ , and  $\tilde{w}_j$  values

	$\tilde{k}_j$			$\tilde{q}_j$			$\tilde{w}_j$		
C <sub>1</sub>	(1.000	1.000	1.000)	(1.000	1.000	1.000)	(0.251	0.328	0.406)
C <sub>2</sub>	(1.083	1.250	1.500)	(0.667	0.800	0.923)	(0.167	0.263	0.375)
C <sub>3</sub>	(1.167	1.417	1.667)	(0.400	0.565	0.791)	(0.100	0.185	0.322)
C <sub>4</sub>	(1.333	1.583	1.833)	(0.218	0.357	0.593)	(0.055	0.117	0.241)
C <sub>5</sub>	(1.500	1.750	1.917)	(0.114	0.204	0.396)	(0.029	0.067	0.161)
C <sub>6</sub>	(1.417	1.667	1.833)	(0.062	0.122	0.279)	(0.016	0.040	0.113)

least important one is Communication skills (C<sub>6</sub>) for the decision-makers. After obtaining the weights of the criteria with fuzzy SWARA method, ESCO alternatives are evaluated with the help of fuzzy MARCOS method. To construct the initial fuzzy matrix, decision-makers evaluate the alternatives under each criterion by using the linguistic variables shown in Table 5.

Evaluation results of decision-makers for alternatives under each criterion are given in Table 6 as linguistic variables, and in Table 7 their equivalents are demonstrated in terms of triangle fuzzy numbers.

Then, aggregated values of decision-makers’ judgments for the alternatives under each criterion are obtained via Eq. 27 (Petrović et al., 2019).

**Table 5** Linguistic variables for alternatives and their response scale (Stankovic et al., 2020)

Linguistic variable	Response Scale
Extremely Poor (EP)	(1, 1, 1)
Very Poor (VP)	(1, 1, 3)
Poor (P)	(1, 3, 3)
Medium Poor (MP)	(3, 3, 5)
Medium (M)	(3, 5, 5)
Medium Good (MG)	(5, 5, 7)
Good (G)	(5, 7, 7)
Very good (VG)	(7, 7, 9)
Extremely Good (EG)	(7, 9, 9)

**Table 6** Evaluation results of decision-makers for alternatives with linguistic variables

		Rating with linguistic variables					
		C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
A <sub>1</sub>	DM <sub>1</sub>	M	G	MG	G	MP	G
	DM <sub>2</sub>	MG	EG	M	MG	M	MG
	DM <sub>3</sub>	M	VG	MP	M	M	G
A <sub>2</sub>	DM <sub>1</sub>	G	MG	MP	VG	VG	EG
	DM <sub>2</sub>	VG	MG	MP	G	EG	VG
	DM <sub>3</sub>	EG	G	M	EG	VG	VG
A <sub>3</sub>	DM <sub>1</sub>	MG	G	MG	MG	G	VG
	DM <sub>2</sub>	M	G	G	G	VG	G
	DM <sub>3</sub>	MP	MG	G	G	G	G

**Table 7** Evaluation results of decision-makers for alternatives with TFNs

		Rating with TFNs					
		C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
A <sub>1</sub>	DM <sub>1</sub>	(3,5,5)	(5,7,7)	(5,5,7)	(5,7,7)	(3,3,5)	(5,7,7)
	DM <sub>2</sub>	(5,5,7)	(7,9,9)	(3,5,5)	(5,5,7)	(3,5,5)	(5,5,7)
	DM <sub>3</sub>	(3,5,5)	(7,7,9)	(3,3,5)	(3,5,5)	(3,5,5)	(5,7,7)
A <sub>2</sub>	DM <sub>1</sub>	(5,7,7)	(5,5,7)	(3,3,5)	(7,7,9)	(7,7,9)	(7,9,9)
	DM <sub>2</sub>	(7,7,9)	(5,5,7)	(3,3,5)	(5,7,7)	(7,9,9)	(7,7,9)
	DM <sub>3</sub>	(7,9,9)	(5,7,7)	(3,5,5)	(7,9,9)	(7,7,9)	(7,7,9)
A <sub>3</sub>	DM <sub>1</sub>	(5,5,7)	(5,7,7)	(5,5,7)	(5,5,7)	(5,7,7)	(7,7,9)
	DM <sub>2</sub>	(3,5,5)	(5,7,7)	(5,7,7)	(5,7,7)	(7,7,9)	(5,7,7)
	DM <sub>3</sub>	(3,3,5)	(5,5,7)	(5,7,7)	(5,7,7)	(5,7,7)	(5,7,7)

**Table 8** Initial fuzzy matrix

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
A <sub>1</sub>	(3, 5, 7)	(5, 7.67, 9)	(3, 4.33,7)	(3, 5.67,7)	(3, 4.33, 5)	(5, 6.33, 7)
A <sub>2</sub>	(5, 7.67, 9)	(5, 5.67, 7)	(3, 3.67, 5)	(5, 7.67, 9)	(7, 7.67, 9)	(7, 7.67, 9)
A <sub>3</sub>	(3, 4.33, 7)	(5, 6.33, 9)	(5, 6.33, 7)	(5, 6.33, 7)	(5, 7, 9)	(5, 7, 9)

**Table 9** Extended initial fuzzy matrix

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
$\tilde{A}(AI)$	(3, 4.33, 7)	(5, 5.67, 7)	(3, 3.67, 5)	(3, 5.67, 7)	(3, 4.33, 5)	(5, 6.33, 7)
A <sub>1</sub>	(3, 5, 7)	(5, 7.67, 9)	(3, 4.33,7)	(3, 5.67,7)	(3, 4.33, 5)	(5, 6.33, 7)
A <sub>2</sub>	(5, 7.67, 9)	(5, 5.67, 7)	(3, 3.67, 5)	(5, 7.67, 9)	(7, 7.67, 9)	(7, 7.67, 9)
A <sub>3</sub>	(3, 4.33, 7)	(5, 6.33, 7)	(5, 6.33, 7)	(5, 6.33, 7)	(5, 7, 9)	(5, 7, 9)
$\tilde{A}(ID)$	(5, 7.67, 9)	(5, 7.67, 9)	(5, 6.33, 7)	(5, 7.67, 9)	(7, 7.67, 9)	(7, 7.67, 9)

$$\tilde{x}_{ij} = (x_{ij}^l, x_{ij}^m, x_{ij}^u) = \left( \min_k(x_{ij}^l), \frac{\sum_{k=1}^K x_{ijk}^m}{K}, \max_k(x_{ij}^u) \right) \quad (27)$$

These aggregated values form the initial fuzzy matrix as it is seen in Table 8.

Later, the extended initial fuzzy matrix is constructed by adding fuzzy ideal  $\tilde{A}(ID)$  and fuzzy anti-ideal  $\tilde{A}(AI)$  solutions with the help of Eqs. 13 and 14. Extended initial fuzzy matrix can be seen in Table 9.

Extended initial fuzzy matrix is normalized by using Eqs. 15 and 16, and then normalized fuzzy matrix is obtained as shown in Table 10.

Weighted normalized fuzzy matrix is obtained via Eq. 17 by multiplying each value in normalized fuzzy matrix with the weights of the criteria. This matrix is shown in Table 11.

Fuzzy  $\tilde{S}_i$  matrix is obtained as given in Table 12 with the help of Eq. 18 by taking the sum of rows of the weighted normalized fuzzy matrix.

Later, the utility degrees of alternatives  $\tilde{K}_i$  are calculated with the help of Eqs. 19 and 20 as it is seen in Table 13.

Fuzzy matrix  $\tilde{T}_i$  is obtained by using Eq. 21 as it is observed in Table 14.

A new fuzzy number  $\tilde{D}$  is obtained by using Eq. 22 as  $\tilde{D} = (0.494, 2.172, 9.539)$ . Later fuzzy number  $\tilde{D}$  is defuzzified via Eq. 23 and  $df_{crisp} = 3.12$  is obtained. Utility functions are determined in relation to the ideal  $f(\tilde{K}_i^+)$  and anti-ideal  $f(\tilde{K}_i^-)$  by using Eqs. 24 and 25 as seen in Table 15.

Later  $\tilde{K}_i^-, \tilde{K}_i^+, f(\tilde{K}_i^-), f(\tilde{K}_i^+)$  values are defuzzified with the help of Eq. 23. These defuzzified values are given in Table 16. The utility functions of alternatives  $f(K_i)$  are calculated by using Eq. 26 and ranking of the alternatives are determined depending on the final values of utility functions.

According to the final values of utility functions, ranking of the alternatives is determined as  $A_2 > A_3 > A_1$ . The textile firm was suggested to choose A<sub>2</sub> ESCO

**Table 10** Normalized fuzzy matrix

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>
$\tilde{A}(AI)$	(0.33, 0.48, 0.78)	(0.56, 0.63, 0.78)	(0.43, 0.52, 0.71)
A <sub>1</sub>	(0.33, 0.56, 0.78)	(0.56, 0.85, 1)	(0.43, 0.62, 1)
A <sub>2</sub>	(0.56, 0.85, 1)	(0.56, 0.63, 0.78)	(0.43, 0.52, 0.71)
A <sub>3</sub>	(0.33, 0.48, 0.78)	(0.56, 0.70, 0.78)	(0.71, 0.90, 1)
$\tilde{A}(ID)$	(0.56, 0.85, 1)	(0.56, 0.85, 1)	(0.71, 0.90, 1)
Weights	(0.251, 0.328, 0.406)	(0.167, 0.263, 0.375)	(0.100, 0.185, 0.322)
	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
$\tilde{A}(AI)$	(0.33, 0.63, 0.78)	(0.33, 0.48, 0.56)	(0.56, 0.7, 0.78)
A <sub>1</sub>	(0.33, 0.63, 0.78)	(0.33, 0.48, 0.56)	(0.56, 0.7, 0.78)
A <sub>2</sub>	(0.56, 0.85, 1)	(0.78, 0.85, 1)	(0.78, 0.85, 1)
A <sub>3</sub>	(0.56, 0.70, 0.78)	(0.56, 0.78, 1)	(0.56, 0.78, 1)
$\tilde{A}(ID)$	(0.56, 0.85, 1)	(0.78, 0.85, 1)	(0.78, 0.85, 1)
Weights	(0.55, 0.117, 0.241)	(0.029, 0.067, 0.161)	(0.016, 0.040, 0.113)

alternative. The firm found the results reasonable and decided to work with A<sub>2</sub> Energy Service Company.

## 6 Conclusions

Energy is a crucial factor of economic activities and has a strategic role for the success of the firms. It is also an important tool for global economic growth as environmental protection gains relevance in nowadays. Due to the limited natural resources, it is important for companies to have sustainable energy policies in the long run.

In recent years, firms work with ESCOs to implement energy efficiency projects and gain competitive advantage by decreasing the energy costs and sustained long-term savings. Therefore, the selection of the right ESCO becomes one of the key success factors. The selection process of ESCO consists of uncertainty and ambiguity, and sometimes the decision-makers can have difficulties in expressing their preferences precisely in crisp values, therefore the evaluations can be expressed in terms of linguistic variables. Fuzzy MCDM methods are appropriate tools to cope with this kind of decision problems. In this paper, the application of a hybrid approach based



**Table 11** Weighted normalized fuzzy matrix

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>
$\tilde{A}(AI)$	(0.084, 0.158, 0.316)	(0.093, 0.166, 0.292)	(0.043, 0.097, 0.230)
A <sub>1</sub>	(0.084, 0.182, 0.316)	(0.093, 0.224, 0.375)	(0.043, 0.115, 0.322)
A <sub>2</sub>	(0.139, 0.279, 0.406)	(0.093, 0.166, 0.292)	(0.043, 0.097, 0.230)
A <sub>3</sub>	(0.084, 0.158, 0.316)	(0.093, 0.185, 0.292)	(0.071, 0.167, 0.322)
$\tilde{A}(ID)$	(0.139, 0.279, 0.406)	(0.093, 0.224, 0.375)	(0.071, 0.167, 0.322)
	C <sub>4</sub>	C <sub>5</sub>	C <sub>6</sub>
$\tilde{A}(AI)$	(0.018, 0.074, 0.187)	(0.010, 0.032, 0.089)	(0.009, 0.028, 0.088)
A <sub>1</sub>	(0.018, 0.074, 0.187)	(0.010, 0.032, 0.089)	(0.009, 0.028, 0.088)
A <sub>2</sub>	(0.031, 0.100, 0.241)	(0.023, 0.057, 0.161)	(0.012, 0.034, 0.113)
A <sub>3</sub>	(0.031, 0.082, 0.187)	(0.016, 0.052, 0.161)	(0.009, 0.031, 0.113)
$\tilde{A}(ID)$	(0.031, 0.100, 0.241)	(0.023, 0.057, 0.161)	(0.012, 0.034, 0.113)

**Table 12** Fuzzy  $\tilde{S}_i$  matrix

	$\tilde{S}_i$
$\tilde{A}(AI)$	(0.256, 0.554, 1.202)
A <sub>1</sub>	(0.256, 0.655, 1.378)
A <sub>2</sub>	(0.341, 0.733, 1.443)
A <sub>3</sub>	(0.303, 0.676, 1.391)
$\tilde{A}(ID)$	(0.369, 0.862, 1.618)

**Table 13**  $\tilde{K}_i^-$  and  $\tilde{K}_i^+$  values

Alternatives	$\tilde{K}_i^-$	$\tilde{K}_i^+$
A <sub>1</sub>	(0.213, 1.181, 5.377)	(0.158, 0.760, 3.731)
A <sub>2</sub>	(0.283, 1.321, 5.631)	(0.211, 0.850, 3.907)
A <sub>3</sub>	(0.252, 1.219, 5.429)	(0.188, 0.784, 3.767)

**Table 14**  $\tilde{T}_i^-$  matrix

Alternatives	$\tilde{T}_i^-$
A <sub>1</sub>	(0.371, 1.941, 9.108)
A <sub>2</sub>	(0.494, 2.172, 9.539)
A <sub>3</sub>	(0.440, 2.003, 9.196)
$\tilde{D}$	(0.494, 2.172, 9.539)

**Table 15**  $f(\tilde{K}_i^+)$ ,  $f(\tilde{K}_i^-)$  values

Alternatives	$f(\tilde{K}_i^+)$	$f(\tilde{K}_i^-)$
A <sub>1</sub>	(0.068, 0.379, 1.723)	(0.051, 0.244, 1.196)
A <sub>2</sub>	(0.091, 0.424, 1.805)	(0.067, 0.273, 1.252)
A <sub>3</sub>	(0.081, 0.391, 1.740)	(0.060, 0.251, 1.207)

**Table 16** Defuzzified values and ranking of alternatives

Alternatives	$K_i^-$	$K_i^+$	$f(K_i^-)$	$f(K_i^+)$	$K_i$	Rank
A <sub>1</sub>	1.719	1.155	0.370	0.551	0.817	3
A <sub>2</sub>	1.867	1.253	0.402	0.598	0.987	1
A <sub>3</sub>	1.760	1.182	0.379	0.564	0.862	2

on fuzzy SWARA and fuzzy MARCOS methods is presented for the selection of the ESCO for a textile firm. The weights of the criteria in the ESCO selection problem are determined with the help of fuzzy SWARA method. This method was preferred for reasons such as ease of calculation, consideration of more than one decision-maker, and being a new method. After determining the weights of the criteria, ESCO alternatives are ranked by using fuzzy MARCOS method. It is a new methodology and proposes a different way for determining utility functions. It also ensures that many criteria and alternatives are considered while maintaining the stability of the method. As a result of the application of the proposed hybrid fuzzy approach, the ranking of ESCO alternatives is obtained as  $A_2 > A_3 > A_1$ . Based on this result, it is advised to the textile firm to select the A<sub>2</sub> ESCO alternative. The top management of the firm has found the results satisfactory and has decided to work with A<sub>2</sub> Energy Service Company.

The proposed approach provides a new insight into the selection process of ESCOs for the firms. Since the developed approach shows a great deal of flexibility, it can be used in other real-life decision problems of firms. Besides, as it is based on expert knowledge and assessments, it can be applied to MCDM decision problems where there is difficulty to access data in terms of crisp numbers. Because, in the hybrid fuzzy approach, the data of the problem is obtained from the decision-makers by using linguistic variables. The contributions of this paper are twofold. First, two relatively new fuzzy MCDM methods are integrated for the first time. Second, the proposed hybrid approach is applied in a different field such as ESCO selection.

In future studies, researchers may focus on the integration of other fuzzy MCDM methods for ESCO selection. Criteria weights can be obtained with other proposed methods in the literature. In addition, the number of the evaluation criteria and the alternatives may be changed according to needs of the firm. Finally, the proposed hybrid approach can be used by firms to find solutions to their other MCDM problems.

## References

- Agarwal S, Kant R, Shankar R (2020) Evaluating solutions to overcome humanitarian supply chain management barriers: A hybrid fuzzy SWARA–Fuzzy WASPAS approach. *Int J Disaster Risk Reduct* 51:101838
- Ilieva G, Yankova T, Hadjieva V, Doneva R, Totkov G (2020) Cloud service selection as a fuzzy multi-criteria problem. *EM Journal* 9(2):484–495
- Kaufmann A, Gupta MM (1988) *Fuzzy mathematical models in engineering and management science*. Elsevier Science Publishers B.V, Amsterdam
- Kaya S, Erginel N (2020) Futuristic airport: a sustainable airport design by integrating hesitant fuzzy SWARA and hesitant fuzzy sustainable quality function deployment. *J Clean Prod* 275:123880
- Kersulienne V, Zavadskas EK, Turskis Z (2010) Selection of rational dispute resolution method by applying new step-wise weight assessment ratio analysis (SWARA). *J Bus Econ Manag* 11(2):243–258
- Mavi RK, Goh M, Zarbakhshnia N (2017) Sustainable third-party reverse logistic provider selection with fuzzy SWARA and fuzzy MOORA in plastic industry. *Int J Adv Manuf Technol* 91:2401–2418
- Mijajlović M, Puška A, Stević Ž, Marinković D, Doljanica D, Jovanović SV, Stojanović I, Beširović J (2020) Determining the competitiveness of spa-centers in order to achieve sustainability using a fuzzy multi-criteria decision-making model. *Sustainability* 12:8584
- Perçin S (2019) An integrated fuzzy SWARA and fuzzy AD approach for outsourcing provider selection. *J Manuf Technol Manag* 30(2):531–552
- Petrović G, Mihajlović J, Čojbašić Ž, Madić M, Marinković D (2019) Comparison of three fuzzy MCDM methods for solving the supplier selection problem. *Facta Univ Ser: Mech Eng* 17(3):455–469
- Stankovic M, Stevic Ž, Das DK, Subotic M, Pamucar D (2020) A new fuzzy MARCOS method for road traffic risk analysis. *Mathematics* 8:457. <https://doi.org/10.3390/math8030457>
- Stevic Ž, Pamucar D, Puška A, Chatterjee P (2020) Sustainable supplier selection in healthcare industries using a new MCDM method: Measurement of alternatives and ranking according to compromise solution (MARCOS). *Comput Ind Eng* 140:106231
- Yazdani M, Alidoosti A, Zavadskas EK (2011) Risk analysis of critical infrastructures using fuzzy COPRAS. *Economic Research-Ekonomiska Istraživanja* 24(4):27–40
- Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353
- Zarbakhshnia N, Soleimani H, Ghaderi H (2018) Sustainable third-party reverse logistics provider evaluation and selection using fuzzy SWARA and developed fuzzy COPRAS in the presence of risk criteria. *Appl Soft Comput* 65:307–319

# Are NBA Players' Salaries in Accordance with Their Performance on Court?



Ioanna Papadaki and Michail Tsagris

**Abstract** Researchers and practitioners ordinarily fit linear models in order to estimate NBA player's salary based on the players' performance on court. On the contrary, we first select the most important determinants or statistics (years of experience in the league, games played, etc.) and utilize them to predict the player salary shares (salaries with regard to the team's payroll) by employing the non-linear Random Forest machine learning algorithm. We are further able to accurately classify whether a player is low or highly paid. Additionally, we avoid the phenomenon of over-fitting observed in most papers by external evaluation of the salary predictions. Based on information collected from three distinct periods, 2017–2019, we identify the important factors that achieve very satisfactory salary predictions and we draw useful conclusions. We conclude that player salary shares exhibit a relatively high (non-linear) accordance with their performance on court.

**Keywords** NBA · Salaries prediction · Variable selection · Non-linear models

## 1 Introduction

Professional athletes' field performance and salaries is a topic that has attracted the interest of numerous researchers (Zimmer and Zimmer 2001; Yilmaz and Chatterjee 2003; Olbrecht 2009; Vincent and Eastman 2009; Wiseman and Chatterjee 2010; Garris and Wilkes 2017). The general question of interest is whether players deserve their salaries as depicted by their performance statistics.

Specifically for the NBA (National Basketball Association), Sigler and Sackley (2000) studied the task of salary prediction using data from the 1997–1998 season but with only three predictor variables, rebounds, assists, and points, per game. Ertug and Castellucci (2013) gathered from the 1989–1990 up to the 2004–2005 period and related the players salaries with a set of predictor variables, most of which were

---

I. Papadaki · M. Tsagris (✉)

Department of Economics, University of Crete, Heraklion, Greece  
e-mail: [mtsagris@uoc.gr](mailto:mtsagris@uoc.gr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_25](https://doi.org/10.1007/978-3-030-85254-2_25)

405

not related to the players' performance on court. More recently, Xiong et al. (2017) performed a similar analysis using more predictor variables measuring the players performance on court for the 2013–2014 season. Sigler and Compton (2018) studied the 2017–2018 season linking the salaries with more predictor variables exposing the players abilities on court.

Based on these (and other) papers, several questions emerge. How can we decide on the predictor variables used to estimate/predict the player salaries? Secondly, are the results obtained from such analyses valid and reliable enough? For instance, can a value of the coefficient of determination as high as 0.6 or 0.7 be a sign of correctness or even suggest that the analysis was successful? We further emphasize that the relationship between the players statistics and their salaries is non-linear and hence linear models are bound to fail in capturing the underlying true association. An additional concern, separate from non-linearity, is model predictability for which internal evaluation has limitations and leads to an over-optimistic performance. These and more matters, discussed later, require delicate treatment which, if not properly addressed, will yield erroneous results.

We deviate from the beaten tracks by first selecting the statistics or determinants with the highest effect on the player salaries. The detection of the important statistics that incorporate the highest amount of information on the players' salaries. We advocate against the use of all available statistics and strongly encourage researchers to apply variable selection algorithms. Not only is it important to determine the appropriate statistics, but it is also crucial for the predictive performance of the models or algorithms applied. We must further decide whether a single set of statistics (*Per game*, *Per 36 min*, etc.) or their combinations contain the highest amount (in linear terms) of information about the salaries and whether feature construction improves our predictions. Do *Advanced* statistics contain more information about the players than the *Per game* or the *Per 36 min* statistics? In either case, do we really require all sets of statistics or a subset of them? Selecting the appropriate statistics, not only removes the noise from the data, but also gives a better insight into the problem.

What can we say about the relationship between the given statistics and the players' salaries? How accurate can our salary predictions be? The second step is to apply sophisticated models to the selected statistics in order to predict the NBA player salaries. Researchers ordinarily apply linear or generalized linear models (e.g., logistic regression), or cluster and discriminant analysis. The downside of these models is their narrow abilities to capture non-linear components when the relationships among the variables are far from linear. Discriminant analysis, for instance, with unequal group covariance matrices, is non-linear but tied to a quadratic function. This gives a higher degree flexibility yet is not flexible enough to capture the associations of interest.

In the next section, we describe the problem of NBA player salary prediction and provide information on the available statistics, how we pre-processed and "cleaned" the data and set the goal of this paper. In Sect. 2, we adumbrate improper approaches that are frequently followed. For instance, employment of linear models or erroneous application of the aforementioned non-linear algorithms to real and simulated data. We describe the tools used for this purpose in the same section; the variable selection

we used to select the appropriate statistics and the machine learning algorithms we employed to predict the player salaries. We further show an example of a false analysis and show that all other methods have fallen in the pitfall of over-fitting. In Sect. 3, we delineate a proper approach depicting how to properly evaluate the models by using the cross-validation (CV) procedure and present the results of our analysis. We further explain why our achieved predictive performance is the highest ever achieved. We finally summarize our findings concluding the paper.

## 1.1 Description of the Data

The starting point of the entire process is to compile all the required information about the players' performance on court, their salaries, the team payrolls as well as other determinants that might prove useful. Our main source of data was [basketball-reference.com](https://www.basketball-reference.com) which is broadly known for providing a great variety of reliable sports statistics. The data acquired were narrowed down to three NBA seasons, 2016–2017, 2017–2018, and 2018–2019. There were available statistics on 486, 540, and 530 players, respectively, for these three seasons.

Throughout the player statistics data accumulated, a multitude of 54 variables<sup>1</sup> provided a plurality of information about each players' performance *Per game*, *Per 36 min*, and *Per 100 possessions*. Those include indexes for field goals, three-point field goals and two-point field goals counted as of total number (FG, 3P, 2P), total number of attempts (FGA, 3PA, 2PA), and percentage of successful attempts (FG%, 3P%, 2P%). The index effective field goal percentage (eFG%), found exclusively on the *Per game* statistics, adjusts for the fact that a three-point field goal is worth one more point than a two-point field goal. In a similar manner, we have free throws (FT), free throw attempts (FTA), and free throw percentage (FT%), offensive rebounds (ORB), defensive rebounds (DRB), and total rebounds (TRB) per game, per 36 min and per 100 possessions. Furthermore, assists (AST), steals (STL), blocks (BLK), turnovers (TOV), personal fouls (PF), and points (PTS) were also included. Among the per 100 team possessions statistics, two more variables were incorporated, offensive rating (ORtg) and defensive rating (DRtg), which are estimates of points produced by players or scored by teams per 100 possessions and of points allowed per 100 possessions, respectively.

In the attempt to obtain a more comprehensive picture of the players' performances, we also consulted the players' *Advanced* statistics, also available on [basketball-reference.com](https://www.basketball-reference.com). This type of statistics displays variables such as Player Efficiency Rating (PER), a measure of per-minute production standardized so that the league average is 15, True Shooting Percentage (TS%), a measure of shooting efficiency

---

<sup>1</sup> There were some common variables though such as their age on February 1 of the season (Age), number of years they have played in NBA (EXP), the team (Tm), and the position (Pos) in which they played. Additionally, the total number of games (G) and minutes (MP) they participated and the number of games they were in the starting five (GS) were displayed on almost all types of statistics.

that takes into account two-point field goals, three-point field goals, and free throws, three-point attempt rate (3PA<sub>r</sub>), which is the percentage of field goals attempts from a three-point range and free throw attempt rate (FT<sub>r</sub>) which indicates the number of free throw attempts per field goal attempt. In addition, the percentage of available offensive rebounds, defensive rebounds, and total rebounds a player took while he was on the court is estimated using offensive rebound percentage (ORB%), defensive rebound percentage (DRB%), and total rebound percentage (TRB%), respectively.

*Advanced* statistics further comprise assist percentage (AST%), an estimate of the percentage of teammate field goals a player assisted, steal percentage (STL%), and block percentage (BLK%), estimates of the percentage of opponent possessions that end with a steal by the player and of opponent two-point field goal attempts blocked by the player, together with usage percentage (USG%), an estimate of the percentage of team plays used by a player and turnover percentage (TOV%), an estimate of turnovers committed per 100 plays. Moreover, we have estimates of the number of wins contributed by a player in total (win shares (WS)), due to his offense (offensive win shares (OWS)), due to his defense (defensive win shares (DWS)), and per 48 min (win shares per 48 min (WS/48)) with the last's league average being approximately 10%. We additionally incorporated the offensive box plus/minus (OBPM), defensive box plus/minus (DBPM), and box plus/minus (BPM), which are box score estimates of the offensive, defensive, and total points per 100 possessions a player contributed above a league-average player translated to an average team. Lastly, the value over replacement player (VORP), a box score estimate of the points per 100 team possessions that a player contributed above a replacement-level (−2.0) player, translated to an average team, and prorated to an 82-game season.

Next on our data collection process, we turned to [espn.com](http://espn.com) and [hoopshype.com](http://hoopshype.com) to attain the fundamental information about the players' income and the 30 NBA teams' payrolls for the seasons under investigation.<sup>2</sup> As far as the players' salaries are concerned, the number of available observations was 594, 598, and 503 for the 2016–2017, 2017–2018, and 2018–2019 season, respectively.

It is of major importance to take into account the teams' payroll when predicting the players' salaries, given the variation of this amount among different teams. During 2016–2017, Utah Jazz's payroll was the minimum across NBA with their contracts summing to \$80 millions, whereas Cleveland Cavaliers spent the highest amount of money, \$130 millions. During the 2017–2018 season, Dallas had the minimum payroll, \$85 millions, whereas Charlotte Hornets had the highest payroll of \$143 millions with Cleveland Cavaliers having second highest, \$137 millions. In our latest season, 2018–2019, Atlanta Hawks had the minimum payroll, whereas Miami Heat had the highest payroll, equal to \$79 millions and \$153 millions, respectively. Markedly, had Miami Heat's best player signed with Atlanta Hawks he would earn

---

<sup>2</sup> [Hoopshype.com](http://Hoopshype.com) provided us with the choice between the absolute nominal value of each team's payroll or the payroll adjusted for inflation based on the current year (from data provided by the U.S. Department of Labor Bureau of Labor Statistics), to which we chose the first for the sake of correspondence between the base years on affiliated monetary values.

around 65% of his Miami salary, and conversely if Atlanta Hawks' best player was traded to Miami he should get 150% times his current salary.

Sigler and Compton (2018) signified that a player's years on the league is a determinant of equal importance for his salary with his performance, as depicted by his statistics. The maximum amount of salary a player can receive is related to the number of years he has played in the NBA and the amount of the salary cap. During the 2017–2018 season, the maximum salary of a player who had at most 6 years of experience was either \$25,500,000 or 25% of the total salary cap, whichever was greater. For a player with 7–9 years of experience, the maximum increased to \$30,600,000 or 30% of the salary cap, and for a player with more than 10 years of experience, his maximum contract could reach \$35,700,000 or 35% of the salary cap. However, a player can sign a contract for 105% percent of his previous contract, even if the new contract is higher than the league limit. Having said this, we made use of [stats.nba.com](https://stats.nba.com) to include each player's experience (number of years in the NBA league) in our inquiry.

## ***1.2 Cleaning and Pre-processing the Data***

The volume of data accumulated needed to be merged into a unified database that would serve the purpose of our analysis. The objective of this process was to associate each player's wage with their statistics, their years of experience, and their team's budget. However, we came across cases of missing information throughout the different sources. For example, there was no available salary or experience for some of the players listed on the statistics database and vice versa. To solve this problem, we solely kept the observations in which we had all three types of information at our disposal. Moreover, some of the players switched teams within the season and, as a result, they were recorded several times on the statistics database, once for every team. In this instance, we preserved the statistics exclusively for the team for which we had information about the salary. On account of better results, it was also deemed necessary to discard all players who participated in less than 10 games during each season, in view of the fact that those observations' contribution to our model's predictability was actually a drawback. Throughout this process of "data cleaning," the remaining observations were 443 players for the 2016–2017 season, 484 players for the 2017–2018 season, and 412 players for the 2018–2019 season. These are considered to be adequate sample sizes.

To make our predictions payroll free, instead of using the nominal wage for each player as the dependent variable on our regression models, we constructed a new variable, the ratio of the player's salary to his team's payroll. This is the players salary share, which will later be used as the dependent variable on our models. The sum of the player salary shares for each team ought to be  $\leq 1$  and in cases it exceeded 1 it was decided to replace the team's payroll with the aggregation of the team's players' salaries at hand.



### 1.3 *Other Possible Determinants of Salaries*

It can be argued that NBA player salaries are not only related to their performance on court but also to publicity and reputation (Ertug and Castellucci 2013). Reputation though is difficult to measure for all players, counting, for example, players' followers in social networks, their contracts with sports companies, promoting activities, etc., and we thus have avoided it.<sup>3</sup> Other contributing factors include player managers that can make hard negotiations with team managers and can achieve higher earnings for their clients. We assert that these factors are projections of the players' image on the court. Highly skilled (and regularly spectacular) players are those who will ordinarily sign contracts with sport companies and will be interviewed more frequently and promoted more, by sports journalists.

A second factor is discrimination, either racial (Kahn and Sherer 1988; Hamilton 1997; Kahn and Shah 2005; Rehnstrom 2009; Wen 2018), nationality-wise (Yang and Lin 2012; Hoffer and Freidel 2014); or exit (Groothuis and Hill 2013). Our personal view is that any alleged discrimination present is fully justifiable by the players' performance. African-American players have better physical skills and are more athletic, which facilitates the quantity of spectacle they offer compared to other players. If those players receive higher salaries simply because they may have better statistics or have a more spectacular type of play, this is by no means evidence of race or country discrimination. Further, foreign players, e.g., Europeans have nourished in a different mentality. American basketball is more athletic than European and usually Europeans require more adjustment time than players drafted from the NCAA. It is perceptible that athletic and physical abilities and the mentality of basketball has caused this alleged discrimination. Investigation of this entails a comparison of the player performance between African-American and white American players and between American and European players, but this is outside the scope of this paper. The same rule applies for the exit discrimination (Groothuis and Hill 2013) who concluded that more athletic players have a higher survival rate in the NBA. We close the discrimination matter by referring to Groothuis and Hill (2013) who used a panel dataset from 1990 to 2008 and failed to find any evidence of either pay or exit discrimination in the NBA.

A third possible determinant factor is TV contracts,<sup>4</sup> which were deemed as not important by Kelly (2017). Kelly (2017) applied a linear regression model where a subset of the TV contracts relevant variables were statistically highly significant, yet the goodness of fit of the model was very poor.<sup>5</sup>

---

<sup>3</sup> For example, to measure a player's level of spectacle we would have to collect the number of dunks, the number of alley-hoops, the number of fake movements, the number of ankle-breaking phases, or any other spectacular movements.

<sup>4</sup> TV contracts contribute to the NBA revenues which determines the salary cap.

<sup>5</sup> This is another case that exemplifies why non-linear models are necessary to yield more accurate predictions.

## 2 Salary Prediction

We will now illustrate some incorrect approaches that are ordinarily followed by researchers. Subsequently, we describe the pipeline for selecting the most important performance determinants that affect the player salaries and how to make the most of the predictive capabilities of those determinants.

### 2.1 Criticism of Some Current Approaches

Sigler and Sackley (2000) related some player statistics (points, rebounds, and assists per game) with the player salaries using a linear regression model and computed an unsurprisingly low coefficient of determination ( $R^2$ ). Ertug and Castellucci (2013) used a linear regression model to estimate the team revenues attributed to ticket sales computing high  $R^2$  values for two models, 0.75 and 0.77. When it came to estimating player salaries, their linear models had a low fit though ( $R^2 = 0.30$  and  $R^2 = 0.31$ ).<sup>6</sup>

The fact that researchers do not select the important determinants prior to fitting the model is an example of what not to do. Ertug and Castellucci (2013), for example, in their seemingly optimal model for the team's ticket-based revenue used 13 variables, out of which only four variables were statistically significant and two of them were highly statistically significant. Retaining the other nine variables in the model does not add but removes value from it in a threefold manner. (a) It is known that the addition of variables in the model leads to higher  $R^2$  values. Thus, the reported value of 0.77 for the  $R^2$  is an over-estimate of the true  $R^2$  of their model. (b) This practice makes the model unnecessarily more complex and (c), in fact, deteriorates the predictive performance of the model. This is associated with the curse of dimensionality (Hastie et al. 2009) and is the main reason why variable selection is necessary, to remove the irrelevant variables that add noise and no information. As an analogue of this task we refer to national teams who select their best among the all-star players when participate in international championships (continental and universal).

Attempting to model the non-linear relationships of the statistics with the player salaries via adoption of a linear model is a policy that should be avoided. A preferable strategy would be to add of square terms in some variables, as this may improve the performance of the model, but perhaps not significantly. Linear models will encapsulate, to some degree, the trend in the variables, but they definitely cannot be used for safe prediction. The following example from basketball suits to convey our message. NBA teams select the most talented rookie players, but solely talent is not enough. It is training that will take those players to the next level and the better the "material" in a team hands the higher its chances to win the championship.

Assessing the goodness of fit of a model via the  $R^2$  is a criticism raising strategy. The performance of a model that has been constructed on some data must be tested

---

<sup>6</sup> In the game of hockey, Vincent and Eastman (2009) applied a quantile regression instead, a robust to outliers regression model, but still linear.

on different data that the model has never “seen.” During training, players test their abilities against one another, but soon they understand each other’s play and perhaps some players perform very well, but only during practice. Players are not getting paid to play well during practice with their teammates, but to play well against new players whose team systems or play they have not seen. Players are always evaluated externally and not internally.

Not all researchers though fall into the aforementioned pitfalls.<sup>7</sup> Wiseman and Chatterjee (2010) performed a variable selection procedure in order to predict the American League Baseball player salaries and also included a quadratic effect in the years of major league service. However, reporting an internal (and hence over-optimistic)  $R^2$  as high as 70.1% was an incorrect decision made by those researchers.

## 2.2 Variable Selection and Prediction Algorithms

To further assist the comprehension of the analysis, we will narrate the LASSO (Least Angle Selection and Shrinkage Operator) variable selection and the Random Forest (RF) algorithm.<sup>8</sup>

### 2.2.1 Least Absolute Shrinkage and Selection Operator

The Pearson correlations between NBA player salary shares and performance measures were deemed statistically significant, but not all of them remain significant when all predictor variables enter a regression model. The LASSO algorithm (Tibshirani 1996) will facilitate the selection of the most important performance measures.

LASSO is a regression model that simultaneously performs variable selection and regularization of the relevant coefficients. It improves the predictive performance by shrinking the regression coefficients so as to reduce over-fitting, and performs variable selection by setting some of them to zero hence discarding variables that are responsible for large variance, therefore making the model more interpretable. LASSO minimizes the following penalized sum of squares:

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|, \quad (1)$$

---

<sup>7</sup> Criticizing all available papers is outside the scope of this paper and hence we do not pursue this further.

<sup>8</sup> We performed the analysis using the open-source software *R* (R Core Team 2020). LASSO is implemented in the *R* package *glmnet* (Friedman et al. 2010), while RF is implemented in *ranger* (Wright and Ziegler 2017).

where  $y_i$  is the  $i$ -th response value,  $x_{ij}$  denotes the  $i$ -th value of the  $j$ -th predictor variable,  $n$  denotes the sample size, and  $p$  is the number of predictor variables. Fine-tuning of the penalty parameter  $\lambda$  is essential since it determines the amount of regularization, the strength of shrinkage, and, ultimately, the number of variables selected for inclusion in the final model. Such is achieved through a cross-validation procedure, where the value  $\lambda$  yielding the lowest estimated prediction error is preferred.

### 2.2.2 Random Forests

The RF algorithm is a fast and flexible data mining approach well suited for high-dimensional data. The algorithm is built upon creating many classification or regression<sup>9</sup> trees. According to Breiman (2001), RF randomly draws a subset of variables and a bootstrap sample<sup>10</sup> and uses only this subset of features to grow a single tree. The process of randomly selecting variables and bootstrap samples is repeated multiple times and the results are aggregated. By creating many random trees (500 or 1000, for instance), one ends up with a random forest.

As stated in the Introduction, the relationship between the player salaries (shares) and their performance on court is not expected to be linear, hence the RF algorithm will allow us to capture the non-linear components of this relationship.

## 2.3 A Note on the Response Variable

We stated earlier that we converted the player salaries into (payroll free) percentages that are on the same basis for everyone. The implications of this transformation to LASSO, which employs a linear regression model, hence a normal distribution, are obvious. Unlike the normal distribution whose support is unbounded, the percentages lie within a restricted range of values. Additionally, predictions with LASSO are not constrained to lie within that plausible range. Not correct specification of a distribution or of a regression that takes into account the space where the response variable is defined is another frequent mistake of researchers and practitioners.

We refrained from using the salary shares in LASSO and transformed the response prior to employing LASSO using the logit transformation  $y^* = \log \frac{y}{1-y}$ , where  $y$  denotes the player salary shares. The logit transformation is well defined when  $y \neq 0$  and  $y \neq 1$ , which holds true in our case. This transformation is not obligatory when RF are used since the predicted values are in fact weighted averages of the observed values (Lin and Jeon 2006).

---

<sup>9</sup> Depending on the nature of the response variable.

<sup>10</sup> Sample with replacement of the same size.

## 2.4 Distribution of the NBA Player Salary Shares

Figure 1 shows the kernel density estimates of the distributions of the salary shares for each season. The differences are rather small and indeed there is no evidence to support that the distributions vary statistically significantly across the three seasons ( $p$ -value = 0.9263).

## 2.5 Internal Evaluation in Our Datasets

We now illustrate the internal evaluation of RF in our datasets and manifest the over-rated performance they seem to be possessing. We standardized our predictor variables prior to the analysis in order to transform all variables into the same scale.<sup>11</sup> We implemented the tenfold CV procedure (described in the next section) to tune the penalty parameter ( $\lambda$ ) of LASSO. We then performed LASSO penalization using the chosen value of  $\lambda$  to select the most important factors which were plugged into the RF algorithm.

We evaluate the predictive performance of RF by contrasting each set of predictions (one set for each hyperparameter) against the true salary shares using the Pearson correlation coefficient (PCC) and the percentage of variance explained (PVE).<sup>12</sup>

$$PCC = \frac{\sum_{i=1}^n (y_i - \bar{y}) (\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (2)$$

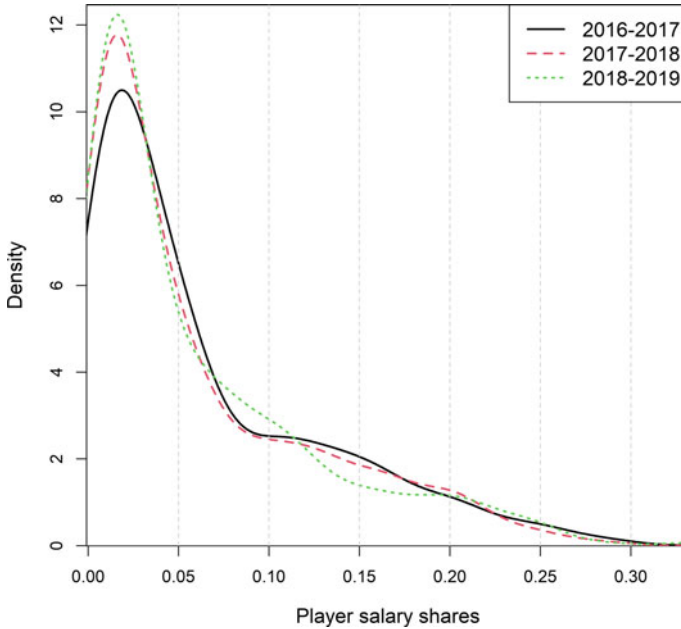
$$PVE = 1 - \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3)$$

where  $\tilde{y}_i$  refers to the predicted value of the  $i$ -th observation and  $\bar{\tilde{y}}$  denotes the mean value. The PVE values for each set of statistics appear in Table 1. We also report the PVE values of LASSO as a comparison of the performance of a linear and of a non-linear algorithm.

Unlike the mean squared error (MSE) or mean absolute error (MAE) the aforementioned metrics have a benchmark value to compare against. The raw values of MAE, or MSE, do not reflect the performance of the model relative to model-free average predictions. On the contrary, for both PCC (2) and PVE (3) their maximum value is 1 indicating excellent predictive performance, whereas the minimal value equal of 0 refers to completely random predictions. Higher values of PCC (2) indicate a higher number of correct model-based predicted orderings, whereas higher values

<sup>11</sup> We explained why this pre-processing step is incorrect in Sect. 3.3.

<sup>12</sup> PVE is equivalent to  $R^2$  for the linear models.



**Fig. 1** Kernel density estimates of the NBA player salary shares across the three seasons

**Table 1** PVE values for each set of statistics and each algorithm across the three seasons

Season	2016–2017		2017–2018		2018–2019	
	RF	LASSO	RF	LASSO	RF	LASSO
<i>Per game</i>	0.910	0.420	0.869	0.246	0.884	0.220
<i>Per 36min</i>	0.901	0.284	0.866	0.223	0.826	0.210
<i>Per 100 possessions</i>	0.900	0.305	0.866	0.223	0.826	0.164
<i>Advanced statistics</i>	0.848	0.163	0.842	0.139	0.862	0.181

of PVE (3) indicate that on average the errors of the model-based predictions are much less than the errors of random, model-free predictions.

We used the PVE<sup>13</sup> (3) for model assessment and perhaps the only safe conclusion we can draw from Table 1 is that non-linear models have superseded the linear model of LASSO. RF always produced PVE values above 0.8 (or 80%) indicating an excellent fit. If we compared these PVE values against the  $R^2$  values reported in previous papers we would be delighted, not only because we outperformed their fit, but also because our PVE values are remarkably high. We will repeat ourselves

<sup>13</sup> The predicted values of LASSO were first back-transformed to percentages using the inverse of the logit transformation,  $y = \frac{1}{1+e^{-y^*}}$ , and  $y$  was used to compute the PVE.

that the cost of this high PVE is interpretability. RF does not produce a coefficient for each predictor variable that could reflect the variable's (marginal) effect on the salary shares.

## 2.6 Model Complexity

The last, but equally important, point to take into account is model complexity. Fitting a highly complex non-linear model does not necessarily yield better prediction. To demonstrate this, we expose below a short script written in R<sup>14</sup> evaluating, internally, a model's performance. We randomly generated a set of 20 predictor variables and a random response variable. We then applied a non-linear model (projection pursuit regression<sup>15</sup>), where each time we increased the complexity of the model and computed the PCC (2) between the observed and fitted values.

```

set.seed(12345)
## generate random predictors
x <- matrix( rnorm(400 * 20), ncol = 20 )
## generate random response
y <- rnorm(400)
pcc <- numeric(10)
for (i in 1:10) {
  ## perform projection pursuit regression
  mod <- ppr(y ~ x, nterms = i)
  ## compute PCC
  pcc[i] <- cor( fitted(mod), y)
}
round(pcc, 3)
0.443 0.575 0.644 0.666 0.748 0.736 0.844 0.933 0.860 0.950

```

Evidently, the PCC between the observed and fitted values increases with model complexity. Further, surprisingly enough, we managed to obtain a high level of correlation when in fact there is no relationship between the response and the predictor variables. This again points out that an internal evaluation draws no safe conclusions as over-fitting occurs. A second source of complexity comes from the fact that we used all available 20 predictor variables and not a subset of them. This is an extra reason why we should have performed variable selection prior to estimating the predictive performance. Penalizing for complexity, e.g., via Bayesian Information

<sup>14</sup> The example is reproducible and will always yield the same results.

<sup>15</sup> We tried this model in our analysis but the results were not that accurate and hence we omitted them.

Criterion could have avoided this phenomenon, but even then, internal evaluation would over-estimate the true performance of the model.

### 3 A More Valid Approach

The previous example indicates how we can get trapped in over-fitting. The reported PVE values refer to the internal evaluation of the models because these are internal PVE values. We will elucidate the correct way to estimate a model's predictive performance (external evaluation) using the  $k$ -fold CV protocol. To obtain an unbiased estimate of the predictive performance we need large sample sizes, a condition we meet because we have information on hundreds of players at each season. Finally, we will demonstrate that the observed performance metrics in Table 1 are actually very high and far from reality.

#### 3.1 The $k$ -Fold CV

The  $k$ -fold CV protocol splits the data into  $k$  mutually exclusive groups, termed folds. The ordinary value of  $k$ , which we also used, is  $k = 10$ , yielding the 10-fold CV. We select one fold and leave it aside to play the role of the test set. The remaining nine folds are combined into what is called the training set. We standardize the predictor variables of the training set only. We then perform variable selection using LASSO and feed the RF algorithm with the selected variables. We use the same selected predictor variables from the test set and we scale them using the means and standard deviations of the same predictor variables from the training set. We use these scaled predictor variables of the test set to predict the values of the response variable (player salary shares) of the test set. For RF we used a range of splits of variables,<sup>16</sup> thus, we end up with multiple predictions, one set of predictions for each hyperparameter, whose predictive performance we compute.

We subsequently select another fold to play the role of the test set and insert the previous fold (previous test set) into the training set and repeat the pre-described pipeline. The process is repeated until all folds have played the role of the test set. In the end, we compute the average predictive performance of RF with each hyperparameter and choose the hyperparameter that yields the highest predictive performance.

---

<sup>16</sup> These are termed hyperparameters and need to be tuned.



### 3.2 *The Essence of CV*

The importance and necessity of any CV protocol can be further appreciated through an investment example. Assume an NBA team manager or team owner who wishes to invest their money on some market, stock exchange, mutual or pension funds, real estate, etc. There are two available investment companies residing in the building right next to his/her. Company A has a long record of remarkably high PVE values in their models. The company gathers the prices spanning from several days ago up to today and fits variable selection and machine learning algorithms and computes the PVE values of the models/algorithms using the same data. It shows no record of predicting future prices though. Company B, on the other hand, applies a different strategy. It again uses historical data, but keeps the old ones for model building and training and treats the most recent ones as the future that must be predicted. The PVE values (of the future predictions) of company B are, perhaps significantly, lower than the PVE values (of the past and present predictions) of company A but are safer predictions of the future. Company A implements the wrong approach described in Sect. 2.1, whereas company B implements the correct strategy described in Sect. 3.1.

### 3.3 *The Importance of Processing the Data in the Training Set*

CV can be seen as a simulation of realistic scenarios. Let us denote the training set by *present* and the test set by *future*. We observe the present and attempt to predict the future. We process (standardize) the data in the training set (present) and use those means and standard deviations to scale the test data (future). Had we standardized the data from the beginning would deviate from the realistic scenario as we would have allowed information from the future to flow into the present. Thus, attempting to transform the data into the same scale prior to performing any CV protocol is erroneous and should be avoided.

But why is standardization so important? Numerous variables listed on the *Per game*, *Per 36 min*, and *Per 100 possessions* refer to percentages therefore deviate between 0 and 1, games played (G) can reach values as high as 82 and players' ages vary between 19 and 42. Furthermore, three-point field goals per 100 team possessions (3P), for example, span between 0 and 7.2 and total rebounds per 100 team possessions (TRB) between 3.0 and 23.8. Likewise, the majority of the *Advanced* statistics are estimates of percentages, while Win Shares (WS) are measured on a scale of  $-1.7$  to  $15.4$  and Box Plus/Minus (BMP) of  $-5.7$  to  $11.1$ , just to name a few. Standardization is a necessary processing strategy in order to prevent our results from being strongly affected by the scale of measurement of the variables.<sup>17</sup>

---

<sup>17</sup> This is the reason why VS algorithms, such as LASSO, require standardized data.

### 3.4 Results of the 10-Fold CV Protocol

Unarguably partitioning the data into 10-folds contains an inherent variability as different partitions will give different results. To robustify our inference against this uncertainty, we repeated the 10-fold CV procedure (variable selection and predictive performance estimation) 50 times and report the aggregated predictive performance of the RF algorithm.

Figure 2 contains the average PCC (2) and PVE (3) for every season using either set of statistics (*Per game*, *Per 36 min*, *Per 100 possessions*, and *Advanced statistics*), along with the corresponding 95% confidence intervals. Overall, use of the *Advanced statistics* resulted in the worst performance among all datasets while the *Per 36 min* and *Per 100 possessions* portrayed a very similar picture, perhaps due to the fact that LASSO was selecting the same statistics. The *Per game* statistics evidently gave the optimal predictions overall with an exception for the season 2017–2018, whose predictions ranked second best with the difference being tiny. The *Per game* statistics also dominated in terms of variance of the predictive performance. The length of the 95% confidence intervals for the true predictive performances are always the shortest, indicating higher stability.

There is a common pattern among the first three sets of statistics. We observe an increase in the predictability as we move from the 2016–2017 to the 2017–2018 season which then decays as we move to the 2018–2019 season. Further, in the last season, we observe the highest variability in the predictive performance and the confidence intervals are the widest observed across the three seasons.

Tables 2 and 3 present the optimal (average) predictive performances of the RF using each set of statistics across the three seasons when LASSO variable selection has been applied prior to RF and when all statistics were fed into the RF. The PCC values are remarkably high, lying in the range of 0.7–0.8. The PVE values are lower,

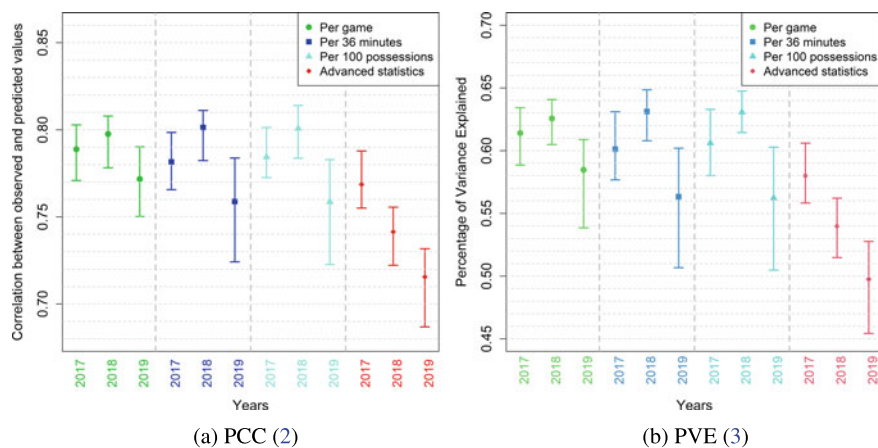


Fig. 2 Predictive performance metrics using each set of statistics across the three seasons

**Table 2** PCC values for each set of statistics across the three seasons

Statistics	With LASSO			Without LASSO		
	2016–2017	2017–2018	2018–2019	2016–2017	2017–2018	2018–2019
<i>Per game</i>	0.789	0.798	0.772	0.783	0.794	0.758
<i>Per 36 min</i>	0.781	0.801	0.759	0.772	0.781	0.750
<i>Per 100 possessions</i>	0.784	0.801	0.759	0.769	0.778	0.747
<i>Advanced</i>	0.769	0.741	0.716	0.742	0.752	0.718

**Table 3** PVE values for each set of statistics across the three seasons

Statistics	With LASSO			Without LASSO		
	2016–2017	2017–2018	2018–2019	2016–2017	2017–2018	2018–2019
<i>Per game</i>	0.614	0.626	0.585	0.600	0.616	0.561
<i>Per 36 min</i>	0.601	0.631	0.563	0.577	0.589	0.540
<i>Per 100 possessions</i>	0.606	0.631	0.562	0.572	0.584	0.534
<i>Advanced</i>	0.580	0.540	0.498	0.534	0.548	0.499

as expected, yet these figures are high in comparison to prior research and most importantly, they were produced by external and not internal evaluation. Further, these two tables clearly visualize the essence of variable selection prior to using RF. The predictive performance changes with and without variable selection changes slightly, but the advantage of LASSO is that it identifies the most important statistics, presented in Table 4.

Table 4 contains the statistics that were most frequently selected by LASSO throughout the 50 repetitions of the 10-fold CV. Overall, experience and minutes played of each player were the two statistics that were always selected regardless of the set of statistics and the year of play. The third statistic was either the games played or the games started, followed by the points, the defensive rebounds, and the field goals attempted. In the Advanced statistics, the USG (an estimate of the percentage of team plays used by a player) and OBPM (box score estimate of the offensive, defensive, and total points *Per 100 possessions* a player contributed above a league-average player translated to an average team). Excluding the *Advanced* statistics, we can see that there seems to be a stability in the selected statistics across the three seasons. In the *Per game*, the last season only substitutes the games played, the field goal attempts, and the defensive rebounds with the points scored. A common feature with the *Per 36 min* is that defensive rebounds do not seem to play a significant role in the last season. When it comes to the *Per 100 possessions*, defensive rebounds never seem to contribute to the salary of the players.

**Table 4** Most important statistics per set of statistics across the three seasons

Statistics	2016–2017	2017–2018	2018–2019
<i>Per game</i>	EXP, MP, G, FGA, DRB	EXP, MP, G, FGA, DRB	EXP, MP, PTS
<i>Per 36min</i>	EXP, MP, GS, DRB, PTS	EXP, MP, GS, DRB, PTS	EXP, MP, GS, PTS
<i>Per 100 possessions</i>	EXP, MP, GS, PTS	EXP, MP, GS, PTS	EXP, MP, GS, PTS
<i>Advanced</i>	EXP, MP, USG, OBPM	EXP, MP, USG, OBPM	EXP, MP, USG, OBPM

The use of *Advanced* statistics did not yield better results than the use of the *Per game* statistics, in fact, the former dataset produced the worst results. We highlight that the PER index is included in the *Advanced* statistics.

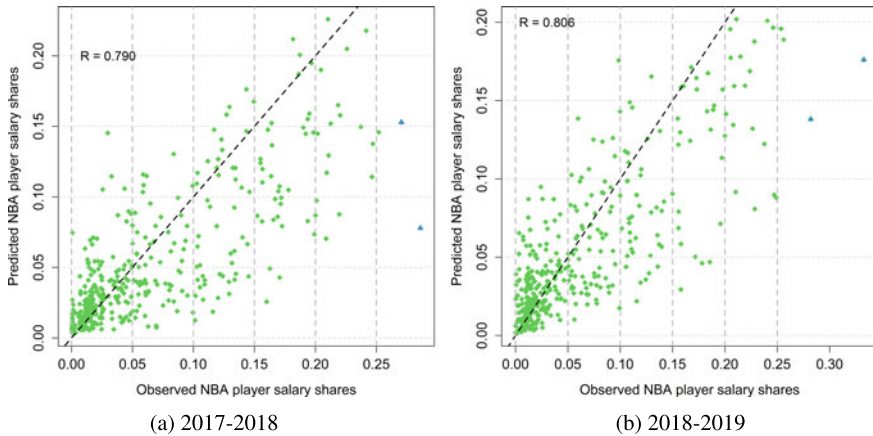
### 3.5 Testing the Predictability of the RF Algorithm

In order to show the validity of the PCC and PVE values reported in Tables 2 and 3, respectively, we used the identified statistics from the 2016–2017 season, fitted an RF, and predicted the salary shares of the 2017–2017 season. We repeated the same task using the statistics from the 2017–2018 season to predict the salary shares of 2018–2019. This way the next season’s data played the role of the validation set, a new dataset that the algorithms never “saw” during the CV protocol. The PVE values were 0.624 and 0.650, respectively, while the PCC values were equal to 0.790 and 0.806, respectively. The observed and the predicted salary shares are displayed in Fig. 3.

### 3.6 NBA Player Salary Share Classes

Let us now provide some in-depth statistics regarding the player salaries. Table 5 shows the distribution of the player salaries across the three seasons.

During the 2016–2017 season, the six highest paid players (in terms of team’s payroll share) were James Harden (Houston Rockets point guard, 29.18%), Al Horford (Boston Celtics center, 28.40%), Russell Westbrook (Oklahoma City Thunder point guard, 26.75%), Kevin Durant (Golden State Warriors power forward, 26.13%), Brook Lopez (Brooklyn Nets center, 25.69%), and Dwyane Wade (Chicago Bulls shooting guard, 25.08%). Among them, James Harden was second in the points per game (29.1), Russell Westbrook and Kevin Durant’s statistics justify their salaries. Surprisingly enough, Al Horford’s statistics do not match his salary, as he was scor-



**Fig. 3** Observed versus predicted player salary shares for 2017–2018 and 2018–2019

**Table 5** Distribution of player salaries across the three seasons

Season	Player salaries in percent of the team’s payroll					
	[0, 5%)	[5, 10%)	[10, 15%)	[15, 20%)	[20, 25%)	[25, 30%]
2016–2017	239	73	48	32	17	6
2017–2018	257	66	49	33	17	3
2018–2019	231	75	39	25	18	4

ing 14 points per game despite playing 32 min. The same is true for brook Lopez and Dwyane Wade whose statistics are rather low.

During the 2017–2018 only three players received more than 25% of the team’s payroll, Paul Millsap (Denver Nuggets power forward, 28.61%), Harrison Barnes (Dallas Mavericks power forward, 27.05%), and Stephen Curry (Golden State Warriors point guard, 25.20%). Stephen Curry was scoring an average of 26.4 points per game, whereas Paul Millsap and Harrison Barnes were as low as 14.6 and 18.9 points per game, despite playing 30 and 34 min per game, respectively.

The four highest paid players (in terms of salary shares) for the last season, 2018–2019, were LeBron James (Los Angeles Lakers small forward, 33.25%), Chris Paul (Houston Rockets point guard, 28.19%), Stephen Curry Golden State Warriors point guard, 25.60%), and Blake Griffin (Detroit Pistons, power forward, 25.36%). Chris Paul was the only one among those 4 to score less than 20 points per game (15.6) even though he was playing 32 min per game. He was giving 8.2 assists per game and stealing the ball 2 times per game, yet these statistics do not match that large salary share.

The aforementioned players were evidently receiving a remarkably high share of their team’s payroll, more than a quarter. LeBron James received an excessively high share, more than a third of Los Angeles Lakers’ payroll, during the 2018–2019

season. We have no evidence to conclude that the highest paid players belong to the champion team. Kevin Durant won the NBA championship with the Golden State Warriors in 2017 and Stephen Curry was a member of the same team that won the championship in 2018. Toronto Raptors won the championship, but their best player<sup>18</sup> is not in the aforementioned list. These players are not the best among NBA and this small piece of information markedly shows that salaries are not always affected by statistics, hence partially explaining why salary prediction is hard to do with only performance statistics.

The blue triangles in Fig. 3a correspond to Harrison Barnes (up) and Paul Millsap (down) indicating that RF predicted correctly that their salary shares should be lower than what they actually received. The blue triangles in Fig. 3b correspond to LeBron James (up) and Chris Paul (down). According to RF, LeBron James was rather over-paid during that year, whereas Chris Paul was evidently over-paid as depicted by his statistics.

### 3.6.1 Salary Share Class Prediction

Table 5 transparently presents that most NBA players receive a small percentage (at most 5%) of the team's payroll. This led us to the second part of our analysis that of discriminating between the low and the higher paid players. To this end, we employed the LASSO and RF algorithms again. In this scenario, LASSO selects the most appropriate statistics by minimizing a more appropriate penalized function

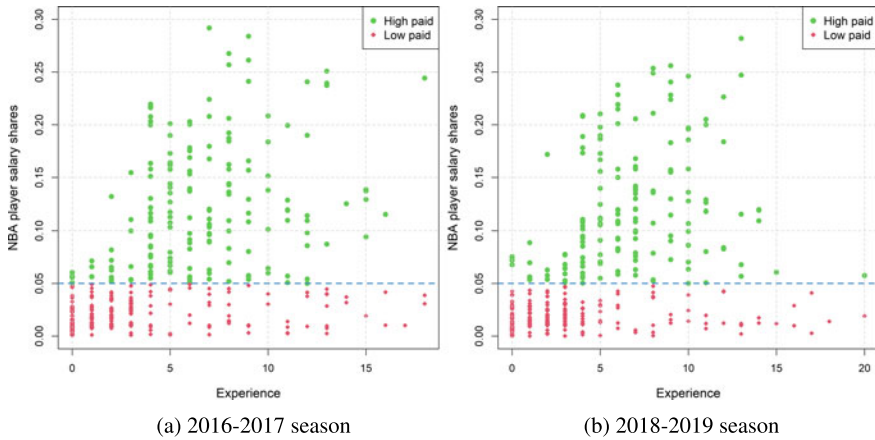
$$\sum_{i=1}^n \left[ C_i \sum_{j=1}^p \beta_j x_{ij} - \log \left( 1 + e^{\sum_{j=1}^p \beta_j x_{ij}} \right) \right] + \lambda \sum_{j=1}^p |\beta_j|, \tag{4}$$

where  $C_i$  takes two values, 0 and 1 corresponding to players receiving lower or more than 5% of their team's payroll, respectively.

Having mentioned earlier that the number of years in the NBA affects the player salaries we visualize their relationship in Fig. 4.<sup>19</sup> Their relationship is clearly non-linear, the Pearson correlations are rather low (0.45 and 0.42, respectively) and there is no apparent threshold to separate the low from the highly paid players. We cannot visually distinguish, in a straightforward manner, the low from the highly paid players. Further, broadly speaking, there is tendency for the salaries to increase, as expected, with three years of service in the league but percentage-wise this is not true for all players.

<sup>18</sup> Kawhi Leonard was receiving 16.78% of Toronto Raptors' payroll.

<sup>19</sup> We present this information for the 2016–2017 and 2018–2019 seasons only, due to space limitations. The scatter plot for the 2017–2018 season, and the scatter plots for the number of games played and the number of games the players were in the starting five were similar and hence omitted.



**Fig. 4** Player salary shares against the number of years in the NBA

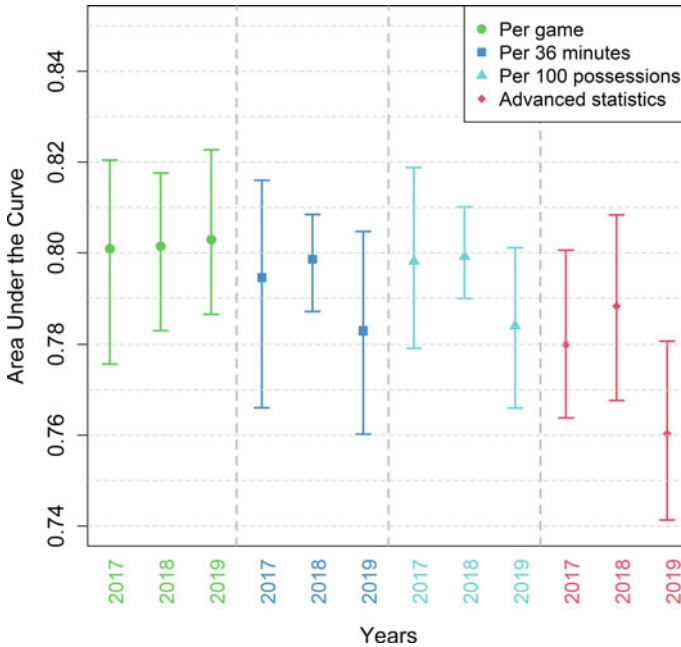
### 3.6.2 Assessment of the Classification Task

We utilized the Area Under the Curve (AUC) to evaluate the classification performance. In this context, AUC shows the probability of correctly classifying a player to the class or group (low or highly paid players) he belongs to. AUC lies within 0 and 1, where 0.5 denotes random assignment. In contrast to the proportion of correctly classified players, AUC is not affected by the distribution of the two groups.

We implemented the same (50 times) repeated 10-fold CV protocol and present the results in Fig. 5. Once again, the *Per game* statistics resulted in the optimal predictive performance, for which the average AUC was always greater than 0.80, whereas the *advanced* statistics yielded the lowest predictive performance. To appreciate the significance of this high value, we can give the following interpretation. The years of experience of a player in the league and the average number of minutes he played for a given season allow to classify him to the low- or high-paid group with a probability equal to 0.8.

In terms of the selected statistics, LASSO was consistently selecting the same statistics as can be seen in Table 6. The number of years the players in the NBA, the average minutes they played in each game, and the number of games they were in the starting five were the most important statistics throughout the datasets and the three seasons.

As a second, validation, step we used the selected statistics from the 2016–2017 season, namely, the number of years in the league and the minutes played, fed them into an RF using the statistics of 2017–2018, and predicted the salary share class of that season. We repeated this task to predict the salary share classes of the 2018–2019 season. The reasoning behind is to test the algorithm’s ability to predict the next season’s salary share classes. The AUC values for the 2017–2018 and the 2018–2019 predictions were 0.811 and 0.841, respectively, corroborating the results of the CV process.



**Fig. 5** AUC using each set of statistics across the three seasons

**Table 6** Most important statistics per set of statistics across the three seasons

Statistics	2016–2017	2017–2018	2018–2019
<i>Per game</i>	EXP, MP	EXP, MP, GS	EXP, MP
<i>Per 36 min</i>	EXP, MP, GS	EXP, GS	EXP, MP, GS
<i>Per 100 possessions</i>	EXP, MP, GS	EXP, GS	EXP, MP, GS
<i>Advanced</i>	EXP, MP, OBPM	EXP, MP, WS	EXP, MP, WS

### 3.7 Further Analysis

We further performed other variable selection algorithms ( $\gamma$ -OMP, Tsagris et al. 2020) and non-linear prediction algorithms such as projection pursuit (Friedman and Stuetzle 1981) and  $k - NN$  (Altman 1992) but their results were sub-optimal and hence omitted. Another strategy was to construct more variables for each dataset, such as square and cubic transformation of each variable, along with all pairwise products of the variables. A second strategy was to combine all variables and the third strategy was to use all variables for each dataset and ignore the variable selection phase. None of these strategies improved the predictive performance of the RF.



## 4 Conclusions

The relationship between NBA player statistics and their salaries (expressed as percentage of the team's payroll) is evidently non-linear and we showed the necessity to apply non-linear models and algorithms. Using real and simulated data we showed the erroneous decisions that can be made when applying linear models. We demonstrated that non-linear models will yield over-optimistic results when they are internally validated. We then described the correct approach to investigate the relationship between a response and many predictor variables and how to correctly estimate a model's predictive performance.

Using the LASSO variable selection we managed to detect the important factors (statistics) that are mostly associated with the NBA player salaries and utilizing the RF non-linear algorithm we predicted the player salaries satisfactorily enough. The level of achieved accuracy is, to the best of our knowledge, the highest ever observed. The validity of the variable selection process and non-linear prediction was evaluated using a repeated cross-validation protocol yielding reliable results.

Predicting NBA player salaries using information on the players' performance on court yields predictions whose accuracy is satisfactory but not as high as one would expect. We argue that key factors mentioned in the manuscript, such as popularity, quantity of spectacle offered, etc. could improve the accuracy of the salary predictions significantly. Another future idea is to switch direction. Instead of investigating the present, whether the players are getting paid according to what they perform on court, one should investigate their future salaries. The level of the contract of a free agent depends not only on his record but also on many factors, such as his age, his playing position, and the available teams among others. For instance, a power forward/center with high performance will sign with a team that is looking for a power forward/center. Additionally, among those teams interested in that player, one must see their salary cap and the players already in that team in order obtain a better picture. Further, we did not include more personal information, such as whether a player is an All-Star, if he is a member of the all-NBA team or the NBA All-Defensive Team, etc. Examination of all those factors could yield more accurate salary predictions than those presented in this paper. Adoption of more complex machine learning algorithms, such as SVM (Drucker et al. 1997) or gradient boosting (Friedman 2001), is another possibility worth exploring.

We close this paper by posing a question. Is it possible that more than one combination of statistics facilitate the prediction of the NBA player salaries? Evidently, the minutes played, the field goals attempted, and the points scored are correlated. By observing the selected statistics in Table 4 we saw that the points scored, substituted the games played, and the field goals attempted, only for the last season. This could be evidence that the variable selection task returns one solution among the many.

**Acknowledgements** We are grateful to Stefanos Fafalios, Nikolaos Pandis, Eleftherios Pavlos, and Zacharias Papadovasilakis for reading earlier versions of this manuscript and providing constructive feedback.

## References

- Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 46:175–185
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: *Advances in neural information processing systems*, pp 155–161
- Ertug G, Castellucci F (2013) Getting what you need: how reputation and status affect team performance, hiring, and salaries in the NBA. *Acad Manag J* 56:407–431
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29:1189–1232
- Friedman JH, Stuetzle W (1981) Projection pursuit regression. *J Am Stat Assoc* 76:817–823
- Garris M, Wilkes B (2017) Socceromics: salaries for World Cup Soccer athletes. *Int J Acad Bus World* 11:103–110
- Groothuis PA, Hill JR (2013) Pay discrimination, exit discrimination or both? Another look at an old issue using NBA data. *J Sports Econ* 14:171–185
- Hamilton BH (1997) Racial discrimination and professional basketball salaries in the 1990s. *Appl Econ* 29:287–296
- Hastie T, Tibshirani R, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media
- Hoffer AJ, Freidel R (2014) Does salary discrimination persist for foreign athletes in the NBA? *Appl Econ Lett* 21:1–5
- Kahn LM, Shah M (2005) Race, compensation and contract length in the NBA: 2001–2002. *Ind Relat: J Econ Soc* 44:444–462
- Kahn LM, Sherer PD (1988) Racial differences in professional basketball players' compensation. *J Law Econ* 6:40–61
- Kelly T (2017) Effects of TV contracts on NBA salaries. Technical report, Department of Economics, Colgate University, USA
- Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. *J Am Stat Assoc* 101:578–590
- Olbrecht A (2009) Do academically deficient scholarship athletes earn higher wages subsequent to graduation? *Econ Educ Rev* 28:611–619
- R Core Team (2020) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria
- Rehnstrom K (2009) Racial salary discrimination in the NBA: 2008–2009. *Major Themes Econ* 11:1–16
- Sigler K, Compton W (2018) NBA players' pay and performance: what counts? *Sport J*
- Sigler KJ, Sackley WH (2000) NBA players: are they paid for performance? *Manag Finance* 26:46–51
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodol)* 58:267–288
- Tsagris M, Papadovasilakis Z, Lakiotaki K, Tsamardinos I (2020) The  $\gamma$ -OMP algorithm for feature selection with application to gene expression data. *IEEE/ACM Trans Comput Biol Bioinform* (accepted for publication)
- Vincent C, Eastman B (2009) Determinants of pay in the NHL: a quantile regression approach. *J Sports Econ* 10:256–277
- Wen R (2018) Does racial discrimination exist within the NBA? An analysis based on salary-per-contribution. *Soc Sci Q* 99:933–944
- Wiseman F, Chatterjee S (2010) Negotiating salaries through quantile regression. *J Quant Anal Sports* 6
- Wright MN, Ziegler A (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 77:1–17

- Xiong R, Greene M, Tanielian V, Ulibarri J (2017) Research on the relationship between salary and performance of professional basketball team (NBA). In: Proceedings of the 8th international conference on E-business, management and economics, pp 55–61
- Yang CH, Lin HY (2012) Is there salary discrimination by nationality in the NBA? Foreign talent or foreign market. *J Sports Econ* 13:53–75
- Yilmaz M, Chatterjee S (2003) Salaries, performance, and owners' goals in major league baseball: a view through data. *J Manag Issues* 15:243–255
- Zimmer MH, Zimmer M (2001) Athletes as entertainers: a comparative study of earnings profiles. *J Sport Soc Issues* 25:202–215

# Financial Distress Prediction Using Support Vector Machines and Logistic Regression



Seyyide Doğan, Deniz Koçak, and Murat Atan

**Abstract** Financial distress and bankruptcies are highly costly and devastating processes for all parts of the economy. Prediction of distress is notable both for the functioning of the general economy and for the firm's partners, investors, and lenders at the micro-level. This study aims to develop an effective prediction model with Support Vector Machine and Logistic Regression Analysis. As the field of the study, 172 firms that are traded in Borsa İstanbul, have been chosen. Besides, two basic prediction methods, LRA was also used as a feature selection method and the results of this model were compared. The empirical results show us, both methods achieve a good prediction model. However, the SVM model in which the feature selection phase is applied shows the best performance.

**Keywords** Financial distress · Support vector machine · Logistic regression analysis

## 1 Introduction

Financial distress, by the simplest definition, is a specific type of financial difficulty that a company faces due to internal or external reasons and tries to overcome. Financial difficulties are the obstacles that the company faces in meeting its obligations. These obstacles are lack of liquidity, lack of owner's equity, failing to pay the debts, and lack of capital (Sun et al., 2014a). Companies face a legally binding bankruptcy if they cannot overcome these obstacles for a long time. Given all these, financial

---

S. Doğan (✉)

Department of Econometrics, Karamanoğlu Mehmetbey University, Karaman, Turkey  
e-mail: [dogans@kmu.edu.tr](mailto:dogans@kmu.edu.tr)

D. Koçak

Department of Econometrics, Osmaniye Korkut Ata University, Osmaniye, Turkey  
e-mail: [deniz.kocakerturk@yok.gov.tr](mailto:deniz.kocakerturk@yok.gov.tr)

M. Atan

Department of Econometrics, Ankara Hacı Bayram Veli University, Ankara, Turkey  
e-mail: [murat.atan@hbv.edu.tr](mailto:murat.atan@hbv.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_26](https://doi.org/10.1007/978-3-030-85254-2_26)

distress can be identified as a long and difficult process that starts with a firm's inability to meet its obligations and extending to bankruptcy.

Classical literature limited the financial failure only to the event of bankruptcy. However, as some authors pointed out, financial failure may not always result in bankruptcy. A company can avoid bankruptcy even in a troubled process by accelerating cash flows through selling assets, downsizing, closure of loss-making transactions (Hashi, 1997). On the other hand, a company can unexpectedly come to the brink of bankruptcy due to unpredictable external shocks such as natural disasters, badly ending cases, global economic and financial crises, even if the company did not face financial difficulties previously (Meyer, 1982). Therefore, a typical commercial distress measurement like bankruptcy cannot refer to financial distress on its own. It is more realistic to assess financial distress as a process, rather than a specific incident, even though it makes it more complex to define and classify exactly. As a process, financial distress corresponds to steps that come sequentially to each other, rather than just one event (Agostini, 2013).

Financial distress can occur in different sizes in companies; while its results can affect an entire economy with a domino effect. The distress of companies can leave states, all stakeholders with whom the firm is connected in the finance and public industries in a difficult situation. Therefore, predicting the distress, by developing a good prediction model will give the company and its stakeholders, creditor institutions an opportunity to decrease the costs that will arise in case of a distress, manage, and monitor the process (Zhou et al., 2015).

*Financial distress prediction* (FDP), which is an important research topic in finance, economy, accounting, and engineering fields, is also called *bankruptcy prediction* and *prediction of company distress*. In general, FDP is the prediction of whether the firm will fail or not based on current financial data of the firms through mathematical, statistical, and artificial intelligence techniques. It is accepted that financial distress often remains under the surface but bankruptcy becomes open and obvious to all upon its declaration, therefore it requires an in-depth analysis (Pindado and Rodrigues, 2005; Doğan, 2020: 13).

In recent years, the academic and industrial interest in this topic has increased because of a growing number of firm bankruptcy with the impact of economic crises. The researchers used classical statistical techniques despite some disadvantages in the first years while they went into the effort of developing early warning models convenient for FDP with the machine learning methods in the recent years. This study used Support Vector Machine, which is a powerful machine learning method. There are many successful FDP studies performed with SVM. This study aimed to contribute to the literature by the selection and parameter optimization phase, whose importance for SVM was recently revealed.

## 2 Theoretical Background

The concept of financial distress emerges as a very important concept in financial research. There are many different solutions for this subject, from univariate ratio analysis to multivariate prediction methods, from traditional statistical methods to artificial intelligence-based machine learning methods, from a single classifier method to hybrid classifier methods designed to combine different classifiers (Sun et al., 2014b; Kumar and Ravi, 2007; Lin et al., 2012). Making financial distress prediction (FDP) through statistical models dates back to the 1960s. The first of those studies was Beaver's (1966) study, which proposed a model with a single variable, and which tried to present the financial distress of an enterprise by dealing with financial ratios individually and thus obtaining a general idea about the financial risk of the enterprise. The study was considered a pioneer study in the finance literature. But in the following years, it was criticized since financial distress or business performance cannot be measured based on a single financial ratio and the prediction capacity would be very low. Following those criticisms, Altman (1968) used statistical methods with multiple variables for the first time through the Z-score model he developed. According to the results of the study, more reliable and consistent findings were obtained by evaluating different financial ratios together with their weights. After Altman's success, some examples such as the multiple-regression analysis introduced by Meyer and Pifer (1970), the logistic regression analysis (LRA) introduced by Ohlson (1980), and the probit model introduced by Zmijewski (1984) were applied in the related field. However, some necessity of traditional methods such as linearity, normality, independent variables of prediction, and the functional form already existing between dependent and independent variables cannot quite be ensured in real-life problems. Today, there are alternative methods, which are less sensitive to the above-mentioned assumptions and which are developed based on artificial intelligence techniques.

Decision Trees (DT) are frequently used in artificial intelligence-based studies carried out on FDP due to their easy understanding and interpretation. Gepp and Kumar (2008), Gepp et al. (2010), and Li et al. (2010) proposed DT, classification and regression trees (CART), C5.0 algorithms for FDP, and showed that they yield better results than multidimensional analysis (MDA). Chen (2011) used the C5.0, CART, and CHAID and LRA methods in his FDP study on businesses registered on the Taiwanese stock exchange. In the findings of the study, it was concluded that the predictive power of decision trees increases even more as the financial distress approaches the year. Genetic programming (GP), which is one of the meta-heuristic methods, was used by Etemadi et al. (2009) for bankruptcy estimation and has been shown to perform better than MDA.

Artificial neural network (ANN) in pattern recognition and classification problems is a highly powerful instrument due to its non-linear non-parametric adaptive learning properties. ANN can very effectively represent and define the non-linear relationship in a data set. ANN was first applied to bankruptcy prediction by Odom and Sharda (1990). They also applied the multiple-variable discriminant analysis (MVDA) to

their sampling of 129 enterprises, 65 of which went bankrupt. As a result, the correct classification rate for MVDA was 74.28%, whereas the rate reached 81.81% for ANN. Many similar studies have emphasized that ANN performs better than statistical methods (Tam, 1991; Tam and Kiang, 1992; Fletcher and Goss, 1993; Zhang et al., 1999; Liang and Wu, 2005).

SVM, developed by Vapnik (1995), has also been of interest to many researchers since they provide considerable results. The most fundamental difference between SVM and ANN is that the structure of SVM is based on structural risk minimization. Because it is aimed to minimize the empirical risk to minimize the training set error in ANN. Conversely, SVM adopts the principle of structural risk minimization, which has been shown to yield better performance than empirical risk, using quadratic programming to predict a single and optimal separator plane in the hidden feature space (Min et al., 2006; Zhongsheng et al., 2007). Fan and Palaniswami (2000), for the first time, applied SVM on three different datasets using the financial ratios suggested by the three models (Altman, 1993; Lincoln, 1984; Ohlson, 1980) best known in the literature. Besides, MDA has tested SVM's success by developing financial failure prediction models with a multi-layer perceptron (MLP) and learning vector quantization (LVQ). Min and Lee (2005) also applied SVM to bankruptcy prediction problems. The results of the study show that when compared to ANN, SVM both gives better results and learning is made possible with a smaller number of training sets. To validate the high classification rate, ANN with backpropagation is compared with multiple-variable discriminant and Logit models, and according to the empirical results, SVM provided better results than all other methods. Shin et al. (2005) compared SVM to ANN to show the effectiveness of SVM, and SVM yielded better empirical results. The study also emphasizes these two important points: first, SVM reaches a better generalization capacity with fewer training sets since it tries to understand the geometric structure of the feature space without reproducing the weights of training samples; second, it makes SVM more useful than ANN, as ANN has certain limitations regarding classification problems. Similarly, Shin et al. (2005) made financial distress predictions for Chinese firms and they compared SVM to the other methods used in the above-mentioned study and reached the same conclusion.

Wu et al. (2007) presented a very comprehensive study in the financial failure estimation study using MDA, logit model, probit model, ANN, SVM methods. In this study, it is aimed to enhance the predictive performance of SVM. For this, researchers have optimized SVM parameters using the Genetic Algorithm (GA). Liang et al. (2016) presented a comprehensive study in which the main classifying method was SVM on 239 successful and 239 failed companies operating in Taiwan Stock Exchange from 1999 to 2009, and SVM inputs were investigated. Machine learning has tested the success of SVM with four methods that have been proven and used in the literature. These methods are k-NN, Naïve Bayes (NB), CART, and MLP. According to the experimental results, SVM has been found as the best prediction model. The reasons why SVM is preferred as a method over other data mining techniques in the present study are that SVM yield equivalent or better results can work with fewer training samples, and has fewer parameters to adjust. For this reason, the main estimation method of the study is determined as SVM. The contribution of the

study to the literature is that it is aimed to try new ways to increase the predictive accuracy rate of SVM. Different processes affect the predictive accuracy of SVM. One of these processes is the determination of the optimal feature set (or variables) that provides quality information to the classifier. The learner may encounter redundant, irrelevant, or interrelated data while understanding the geometric structure of the classifier property space. When too much and unnecessary information is given to the model as input, a lot of time and cost will be spent and even the model's suitability rate will decrease slightly (Piramuthu, 2004; Huang and Wang, 2006). However, it is not an easy way to interpret or exclude unnecessary information. For this reason, it is an important issue to filter large amounts of data and intensify it to provide more information especially in financial failure estimation (Tsai, 2008). In most of the current studies, the financial ratios that provide information were chosen from the financial ratios produced by the prediction models previously made. The classification ability of these models will largely depend on studies in which the selected financial ratios are taken (Wu et al., 2006).

In the first studies in the FDP literature (Beaver, 1966; Altman, 1968), the feature selection process was generally carried out using a qualitative approach such as the popularity of features (financial ratios), good results in past studies, or based on expert opinion. This approach has been replaced by quantitative selection techniques over time. Jo et al. (1997), Atiya (2001), Park and Han (2002), Shin and Lee (2002), Min and Lee (2005), Ding et al. (2008), Chen (2011), Li and Sun (2012) selected the features using statistical methods such as progressive regression, t-test, and correlation matrix, factor, and principal component analysis, which are examined in the filter methods category. Min et al. (2006) and Wu et al. (2007) preferred GAs examined in the wrapper feature selection methods category. In these studies, it was emphasized that the power of the prediction model depends on the selected prediction method and feature set. However, another important situation that increases the estimation performance is to investigate the optimal parameter set. SVM has two important parameters called "C" and "gamma". There are lots of studies that emphasized the parameter optimization improves the performance of SVM (Wu et al., 2007; Shin et al., 2005). But there are a limited number of studies investigating both the optimal parameter pair and the optimal feature set. In this study, both parameter optimization and feature selection methods are used for SVM. The preferred feature selection method in the study is LRA. Despite some limitations, LRA is a multivariate statistical method that is frequently preferred in the studies of financial failure estimation. For this reason, it will be used as an alternative method in the study to test the success of SVM. For parameter optimization, the Grid search technique, which is one of the easy and effective methods, is preferred and this technique is presented in Sect. 3. In the next section, the empirical results are summarized. The final section presents a general summary of the study.



### 3 Proposed Methods for the Prediction Model

This chapter presents the working principle of SVM for a typical two-class classification problem and explains the LRA, which is a multiple-variable statistical technique. For detailed explanations about SVM, please refer to Gunn (1998), Smola and Schölkopf (1997), and Cristianini and Shawe-Taylor (2000).

#### 3.1 SVM Classifier

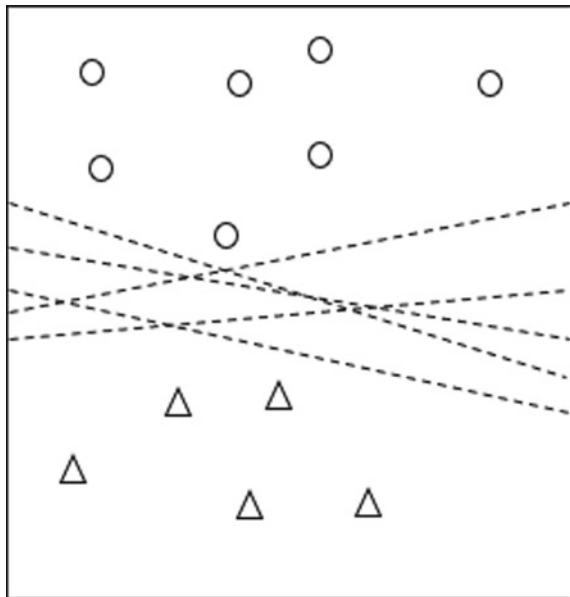
The sample-class labels pair,  $(x_i, y_i)$ ,  $i = 1, 2, \dots, m$ ,  $x_i \in \mathfrak{R}^n$ , and  $y_i \in \{+1, -1\}$  which has  $p$  number of feature (attributes) and which comprises the training set as the linear hyperplane which will separate S training set to represent the class that output samples represent is formulated as follows:

$$w \cdot x + b \quad (1)$$

There can be many linear planes that separate the problem linearly. This can be seen in Fig. 1:

However, it is aimed to find the most suitable separator hyperplane. This hyperplane maximizes the distance between support vectors from different classes, which is called the margin. The distance between  $\langle w, x \rangle + b = 0$  separator plane and the

**Fig. 1** Linear classification



newly observed  $x'$  pattern is determined by  $|\langle w, x' \rangle + b|/\|w\|$ . Each training pattern is at least  $\Delta$  distant from decision boundary and the distance of each training sample from the hyperplane for  $y_i \in \{+1, -1\}$  is determined on condition that

$$\frac{y_i[\langle w, x_i \rangle + b]}{\|w\|} \geq \Delta, \quad i = 1, \dots, n \tag{2}$$

by the equality in the limit value Eq. (3).

$$\frac{1}{\|w\|} (\min_{x:y=1} |\langle w, x_i \rangle + b| + \min_{x:y=-1} |\langle w, x_i \rangle + b|) = \frac{2}{\|w\|} \tag{3}$$

The hyperplane that best separates the training samples is the plane that minimizes the equation  $\eta(w) = \frac{1}{2}\|w\|^2$ . Finding the optimum hyperplane for separable data is a quadratic optimization problem defined by linear limits. The problem is modeled as follows:

$$\begin{aligned} & \underset{w,b}{\text{Min}} \frac{1}{2} w^T w \\ & \text{subject to : } y_i(\langle w \cdot x_i \rangle + b) - 1 \geq 0 \end{aligned} \tag{4}$$

If the problem has a very large data space, then it is not practical to look for a solution through the primal model. Therefore, it will be beneficial to construct the dual of the problem. For that Khun-Tucker theorem is used (Srang 1986: 538–540) and it is of two steps. In the first step, an unrestricted optimization problem is formed using the Lagrange function:

$$L_D(w, b, \alpha) = \frac{1}{2} w^T \cdot w - \sum_{i=1}^m \alpha_i y_i (\langle w \cdot x_i \rangle + b) - 1 \tag{5}$$

In the above-mentioned equation,  $\alpha_i$  is the dual Lagrange multipliers and this multipliers should be maximized by the condition,  $\alpha_i \geq 0$ . On the other hand, when  $w$  and  $b$  are taken into consideration, the Lagrange function should be minimized. Therefore, the optimal value point of the Lagrange function is required. When Karush Khun-Tucker (KKT) conditions are to be satisfied in order to find the derivation of the function according to  $w$  and  $b$ , and to express it only according to  $\alpha_i$  parameter, the restricted optimum function is rewritten. That is the second step of forming the dual model. To form the dual model, the Lagrange function is rearranged using KKT conditions. Thus, the formulation of the dual problem is determined by:

$$\begin{aligned}
 \text{Max}_{\alpha} L_D(\alpha) &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
 \text{subject to: } & \sum_{i=1}^l \alpha_i y_i = 0 \quad , \quad \alpha_i \geq 0, \quad i = 1, \dots, m
 \end{aligned}
 \tag{6}$$

The Lagrange function should be maximized based on the non-negative variable  $\alpha_i$  with the aim of finding the optimal separator hyperplane. In the dual optimization problem, the  $w^*$  and  $b^*$  hyperplane parameters determine  $\alpha_i$ . Thus, the optimal separator decision function  $f(x) = \text{sgn}(\langle w^* \cdot x \rangle + b^*)$  is rewritten:

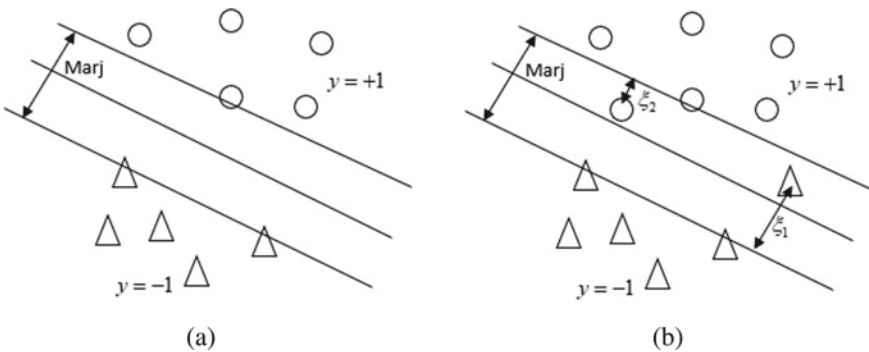
$$f(x) = y = \text{sig} \left( \sum_{i=1}^m y_i \alpha_i^* \langle x_i, x \rangle + b^* \right)
 \tag{7}$$

In a typical classification problem,  $\alpha_i$  smallest sub-set of the Lagrange multipliers tends to be larger than zero. Besides, these non-negative training vectors are geometrically very close to the optimal separator plane. These vectors are termed support vectors and the optimal separator hyperplane is defined only on these support vectors.

If the problem is complex and non-linear, the margin could have a negative value and the appropriate solution area of the problem is empty. In order to overcome this situation, which makes the solution impossible, either you need to relax the strict inequalities, which is called ‘‘soft margin optimization’’, or the problem is made linear using kernel trick. The soft margin optimization can be applied to make a small change in the solution explained above for linearly inseparable data.

In Fig. 2 below, (a) is an example to data that is linearly separated by the maximal margin, and (b) is an example to data that cannot be separated linearly.

In the second situation, the data can be linearly separated by assuming that a specific error is assigned for misclassified samples. In this case, the problem aims to find the hyperplane that minimizes the training errors by means of slack variables:



**Fig. 2** Linearly separable data (a), and Linearly inseparable data (b)

$$\begin{aligned}
 & \underset{w,b,\xi}{Min} \quad \frac{1}{2}w^T w + C \sum_{i=1}^m \xi_i \\
 & \text{subject to : } y_i((w \cdot x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i > 0, \quad i = 1, \dots, l
 \end{aligned} \tag{8}$$

In the above-mentioned model, the penalty parameter on training errors is represented by C, and the non-negative slack variable is represented by  $\xi_i$ . This optimization problem can be solved via the Lagrange multipliers technique. The solution of problem is furthered almost in the same way as in the linear learning case. The Dual model is given below:

$$\begin{aligned}
 & \underset{\alpha}{Max} \quad L_d(\alpha) = \sum_{i=1}^m \alpha_i + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\
 & \text{subject to : } \sum_{i=1}^n y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m
 \end{aligned} \tag{9}$$

In model (9), the majorant of the Lagrange variable is represented by penalty parameter, C, and this parameter is predetermined by the user. Besides, the optimal separator hyperplane function is the same as Eq. (7). The mapping function  $\phi$  is applied for training samples in the non-linear SVM. Using the appropriate kernel function defines dot product (inner product) in feature space, the classifier could separate non-linear data. The Kernel function given in Eq. (10) uses the space of the inner product that we have used in the objective function in the Dual model (9).

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{10}$$

$$\begin{aligned}
 & \underset{\alpha}{Max} \quad L_d(\alpha) = \sum_{i=1}^m \alpha_i + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\
 & \text{subject to : } \sum_{i=1}^n y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m
 \end{aligned} \tag{11}$$

When we follow the solution stage in the linearly separable case, the decision function is derived from  $f(x) = y = sig\left(\sum_{i=1}^m \alpha_i^* y_i (K(x_i, x_j) + b^*)\right)$ . Besides, it must be said that there are lots of kernel functions that enhanced SVM to get the optimal result. The most commonly used of those functions are polynomial (12), radial basis (13), and sigmoid (14) kernels (Burges, 1998; Liao et al., 2004).

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d \tag{12}$$

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{13}$$

$$K(x_i, x_j) = \tanh(K(x_i \times x_j) - d) \quad (14)$$

### 3.2 Logistic Regression Analysis

Logistic regression is a regression analysis used to predict a dependent variable with two categories. The categories of the dependent variable here are formed by using a coding scheme as zero or one to signify that an event has occurred or has not occurred. LRA aims to find the most appropriate model to determine the relationship between a two-category dependent variable and a number of independent variables (Caesarendra, Widodo and Yang, 2010). In this manner, the logistic function with  $p$  number of independent variables is expressed as in (15):

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} \quad (15)$$

where, the statement  $P(Y = 1)$  represents the probability of the relevant event of the dependent variable to occur, whereas,  $\beta_0, \beta_1, \dots, \beta_p$  represent regression coefficients. In the case that the dependent variable represents the probability of the relevant event to occur, the output variables comprise of a number of responses restricted between 0 and 1. Logistic regression also provides a linear model, the natural logarithm of the rate of  $P(Y = 1)$  to  $1 - P(Y = 1)$  in the logistic regression model:

$$g(x) = \ln\left(\frac{P(Y = 1)}{(1 - P(Y = 1))}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (16)$$

$g(x)$  in the Eq. (16) has several features desired in a linear regression model. The independent variables here can be integrated in the model as a combination of continuous and categorical variables. In the analysis, to predict  $\beta_0, \beta_1, \dots, \beta_p$  parameters, the maximum probability prediction is applied after the transformation of the dependent variable to logit variable (Dreiseitl and Ohno-Machado, 2002; Kurt, Ture and Kurum, 2008; Yilmaz, 2009).

## 4 Experimental Study

In the SVM literature, many different model suggestions have been made within the scope of testing and strengthening the success of the method. One of these models is LRA, which is one of the multi-variable statistical techniques. The results of the analysis, which we call the logit model, have been compared to the results

obtained by SVM. In another model, the logit model is used as a feature selection technique and with the variables which have been found significant and which would increase its prediction performance, another analysis was done by SVM. The obtained results from the proposed models have been discussed, and the comparisons are visualized through graphs. In this study, developed SVM model has been designed via MATLAB 9.4 (R2018a)—The Language of Technical Computing program and LIBSVM software system (Chang and Lin, 2011). Besides, the IBM SPSS Statistic-21 package program has been used for LRA.

## 4.1 Datasets

The firms that will be used for financial distress prediction operate in the manufacturing industry and sub-sectors of this industry. Besides, these firms are traded on the BIST stock exchange. Within the scope of these given, 172 of the firms constitute the datasets of the research. Considering that the firms which are subject to Capital Market Law (CML) and traded in Borsa Istanbul (BIST or Stock Market) have prepared their financial statements in accordance with the international financial reporting standards since 2007, the period between 2010 and 2017 has been determined as the “Research Period”. Besides, 24 financial ratios in 6 groups were used in the research. These ratios have been obtained from the firms’ annual balance sheets which are updated through footnotes. Using financial ratios makes it possible to control any potential problem that might occur due to the size of the enterprise and sector differences, and to minimize the impacts of those factors. Therefore, financial ratios, which are frequently used and considered important for firm distress predictions in the literature and which are statistically effective predictors, have been preferred. The financial ratios are given in Table 1. The balance sheets and income statements of the firms whose shares are traded in the Stock Market during the whole or part of the Investigation Period have been obtained by using Finnet Analysis Program.<sup>1</sup>

The “success” or “distress” situations of the firms were used as classifying variables in this research. Based on the definitions regarding the concepts of financial distress in the literature reviewed within the framework of the study, the financial distress criteria have been determined. According to Beaver (1966), Deakin (1972), Aktaş (1993), Altman, Zhang and Yen (2007), Özdemir (2011), these criteria are as follows:

1. That the enterprise has filed for bankruptcy or has gone bankrupt,
2. That the enterprise has made a loss in the last 3 years,
3. That the enterprise has been delisted from stock exchange,
4. That the enterprise has a negative equity,
5. That the enterprise has been on the watchlist firms market for over a year,

---

<sup>1</sup> Finnet: Financial Information News Network. Web: <https://www.finnet.com.tr/FinnetStore/Tr/Urun/Fta40>.

**Table 1** Financial ratios

Definitions	Codes	Financial ratios
Growth rates	X1	Asset Growth (%)
	X2	Share Equity Growth (%)
	X3	Net Sales Growth (%) (Annual)
Valuation ratios	X4	Market Value / Net Sales
	X5	Market Value / Book Value
Operating ratios	X6	Accounts Receivable Turnover (Annual)
	X7	Stock Turnover (Annual)
	X8	Asset Turnover (Annual)
Financial structure ratios	X9	Fixed Assets / Assets
	X10	Short Term Loans / Share Equity
	X11	Short Term Loans / Assets
	X12	Share Equity / Assets
	X13	Short Term Loans / Total Loans
	X14	Share Equity / Real Assets
	X15	Loan Capital Ratio (%)
	X16	Total Loan Growth (%)
Profitability ratios	X17	Net Profit Margin (Annual)
	X18	Return on Assets (%) (Annual)
	X19	Real Operating Profit Margin (Annual)
	X20	Profit Capital (%) (Annual)
	X21	Gross Real Operating Profit Margin (Annual)
Liquidity ratios	X22	Quick Ratio
	X23	Current Ratio
	X24	Current Assets / Total Assets

6. That the enterprise has lost 10% of its total assets, and
7. That the enterprise has restructured its debts.

The enterprises that comply with at least one of the above criteria have been considered “distressed”, and all of those that do not as “non-distressed”. The distressed or non-distressed situations of all 172 firms in our data set have been identified. There are firms that were distressed all through the sampling period or firms which suffered financial distress for only one year and were non-distressed for the rest of the years. The exact opposite situation is also available. Many FDP researchers have used a balanced sample in which class frequencies are distributed as 50–50% (Altman, 1968; Park and Han, 2002; Shin et al., 2005; Sun and Li, 2011). However, most real-life problems have unbalanced class distribution (Liu et al., 2009). According to Zmijewski (1984), if the proportions of distressed and non-distressed classes differ clearly from the real-world stack, the prediction ability of the model is distorted. So

the choice covers the whole spectrum in order to avoid any selection bias, firms have been randomly selected with their financial ratios for the years in question and added to the sampling. In the entire data, it was observed that 71 of the firms are classified as distressed firms, and 101 of the firms as successful firms. It was divided into two groups. Since there is a consensus in the literature, the data set has been randomly split into two: training and testing set (80–20%).

## 4.2 Study Design and Experiments

The outline of the process that has been proposed for the application part of the study is presented in Fig. 3. The detailed explanations are as follows:

### 4.2.1 Kernel Function

Different kernel functions promote SVM in finding the optimal result. Also, it is possible for the user to write their own kernel function based on the structure of the problem. The polynomial, radial basis, and sigmoid kernel are the most used kernel functions (Liao et al., 2004). Since Radial basis function (RBF) can classify multidimensional data, it is the most widely used kernel. When compared to the polynomial kernel, it is known that RBF has fewer parameters. In several studies, RBF is compared to other kernel functions and no significant difference is observed.

In this study, the radial-based kernel function is used. Because RBF for SVM has been accepted as an effective choice in finding the most suitable result.

There are two significant parameters used in SVM that are called C and gamma. The selection of the value of C, which is called the penalty parameter, affects the classification output. If we assign a very high value to C, the classification accuracy rate during the training will be very high. However, the accepted model will most probably have a very low accuracy rate on the test data. If we select C to be very small, it is known that the classification accuracy rate will not be satisfactory. Therefore, the model is impractical. Gamma parameter, on the other hand, has a higher impact on the classification output than does C, because the value of gamma affects the separation output in the feature space. Assigning very high values to gamma leads to over-fitting and very low values to under-fitting (Pardo and Sberveglieri, 2005).

### 4.2.2 Parameter Optimization

The easiest way to adjust C and gamma parameter is the Grid search technique (Hsu et al., 2003). In this technique, the identification of the appropriate parameter to ensure a high classification accuracy rate is done by trying all different combinations between the lower limit and the upper limit determined for gamma and C. As can be seen in Fig. 4, the limits for C range from  $2^{-5}$  to  $2^{15}$ . Besides, the limits for gamma



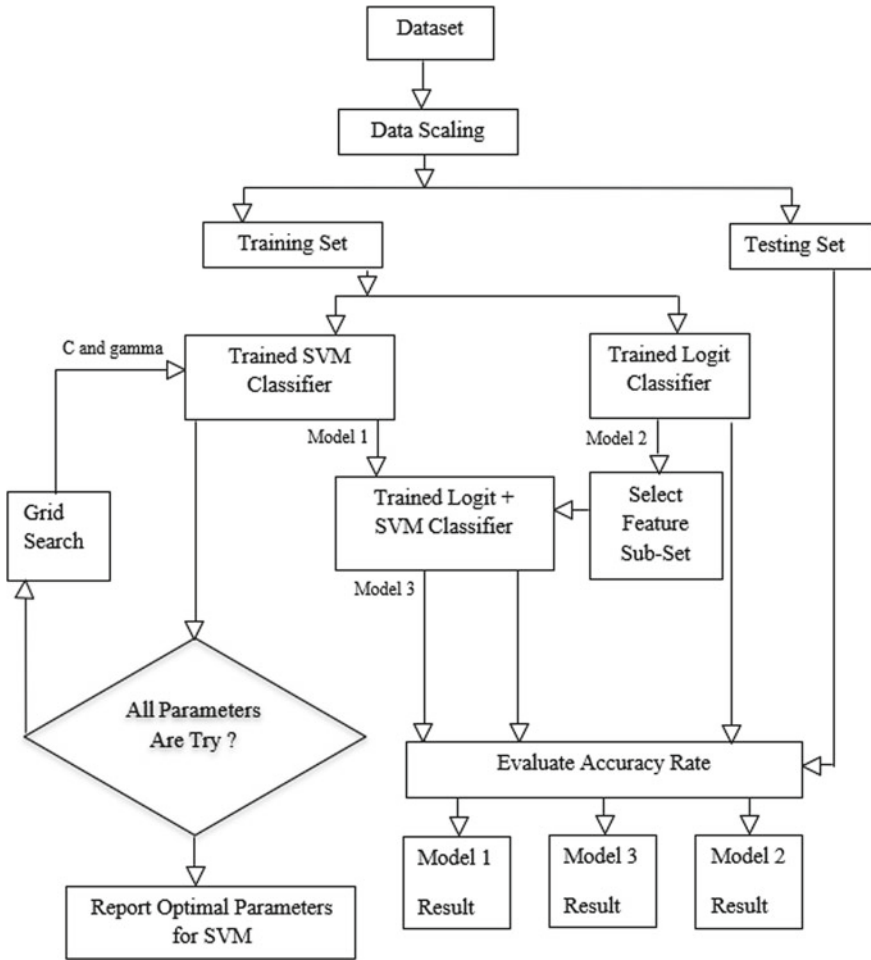
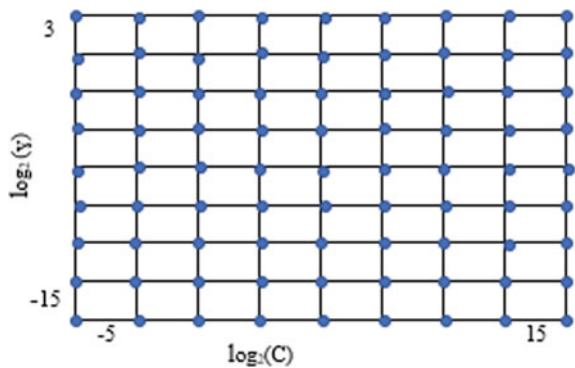


Fig. 3 The proposed analysis process for financial distress prediction

Fig. 4 Grid search



range from  $2^{-15}$  to  $2^3$ . Here, 110 different results are tried and the cross-validation rate for each parameter is calculated. Then SVM training process is initiated with the parameter pair that yields the best cross-validation rate.

In this technique, which is a local search technique, the interval determined for the parameter values should be well adjusted (Lin et al., 2008). A very wide interval means wasted calculation time and determining a narrow interval might indicate that the satisfactory results are left out of the search space, or in other words, that good results are sacrificed. Determining an appropriate parameter for SVM is a separate area of study in itself and it is yet to be developed.

### 4.2.3 Feature Selection

The accuracy rate of SVM is not only affected by C and gamma parameters; the quality of the data set also effect this rate. For instance, a high correlation between features influences the solution results. Excluding an important feature from the model may reduce the accuracy rate. Conversely, some features included in the data set may not affect results or may contain noise.

Feature selection methods are analyzed under three categories as filter and wrapper (Liu and Motoda, 1998), and embedded (Saeys et al., 2007). As filter methods, factor analysis (FA), the principal components analysis (PCA), independent components analysis (ICA), and discriminant analysis (DA) are mostly used. As for wrapper methods, mostly meta-intuitive techniques with a road map which are based on the exploration of the optimal sub-set are used. In embedded techniques, random forest walk, the vector weights of SVM, and logistic model weights are used. Filter methods are fast, but they do not guarantee to give the optimal sub-set; wrapper methods work slowly and give the best *approximate* optimal solution. Embedded methods require more complicated calculations than wrapper methods since they work interactively with the classifier. While the outputs of the filter and wrapper methods are estimators, in embedded methods the output is an estimator and a feature sub-set. Based on Min and Lee (2005), LRA was used in the feature selection phase in the present study.

### 4.2.4 Data Pre-processing

Data pre-processing is applied not to have numeric difficulty during calculations and also to ensure that the large values of the variables are not affected by small values. Moreover, pre-processing appears to be a requirement for many machine learning techniques. The raw data is transformed using the formula given in Eq. (17).

$$Z_{score} = \frac{X_i - X_{mean}}{S} \quad (17)$$

where  $X_i$  is the raw value that each variable takes,  $X_{mean}$  is the average of variable values, and  $S$  is standard deviation. Thus, raw financial ratios are normalized, with their average as zero and standard deviation as unit across samples.

#### 4.2.5 Cross Validation ( $k$ -fold)

In order to make sure that we have developed a model that would assign the newly added data in the sample to the correct class, the model must have an acceptable accuracy rate on the test data set which was kept out of the analysis independently. The most reliable way to do so is to divide the data into  $k$  parts and to keep each time 1 part aside independently as the test set, and then train the model on the remaining  $k-1$  parts. This method is called cross-validation. The advantage of cross-validation is that the test data set kept aside for each time is independent and increases the reliability of the results (Huang and Wang, 2006).  $k$ -fold cross-validation method was first applied in Salzberg’s study in 1997 taking  $k = 10$  (Salzberg, 1997).

The parameters of the method we are going to use in the application stage are optimized by the Grid search technique. The parameter pairs, and therefore, the conformity rates will change in each iteration. For that reason, in the evaluation of prediction results,  $k$ -fold ( $k = 10$ ) cross-validation rate is taken into consideration.

#### 4.2.6 Performance Evaluation

The confusion matrix is used with the aim of comparing the predictions of the model with actual results. The  $2 \times 2$  confusion matrix to be used for a two-class example is presented in Table 2. On the left column of the table are the estimated class values of the samples kept aside as the test data set, and on the upper line are the actual class values.

In some cases, an example in the positive class might also be classified as positive in the prediction, which is called true positive (TP) separation; on the other hand, it is also possible that an example in a positive class might have been predicted to be placed in a negative class (false negative (FN) separation), which is called Type 2 error. In the exact opposite case, an example in a negative class might have been predicted to be in a negative class (true negative (TN) separation), or in a positive class (false positive (FP) separation). This is an indication of Type 1 error. The sensitivity which is called the true positive rate and specificity which is called true negative rate provides significant information about how the classifier separates the positive and negative limits. To

**Table 2** Confusion Matrix

		Actual	
		(+)	(-)
Prediction	(+)	(TP)	(FP)
	(-)	(FN)	(TN)

evaluate the performances of the models, some performances criteria in the related literature are used criteria. The formulas of these performance criteria (accuracy, sensitivity, specificity, certainty, and Matthews correlation coefficient (MCC)) are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (18)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (19)$$

$$Specificity = \frac{TN}{TN + FP} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (22)$$

#### 4.2.7 Model Propositions

In order to obtain a powerful and useful prediction model, three different models have been proposed. Explanations about the models are presented under the titles below; the results and interpretations are discussed in the sect. 4.3 “*Empirical Results and Discussion*”.

**Model 1:** The Analysis by the Support Vector Machines. In Model 1, all variables (Table 1) are used. These variables are the financial ratios which are most commonly encountered in the literature and which provide in many studies significant information regarding explaining financial distress. For the dependent variables of the sampling of 172 firms, only SVM, the support vector machine, the parameters of which have been optimized, has been applied in Model 1. This model has been named Grid SVM.

**Model 2:** The Analysis by the Logistic Regression. In Model 2, all variables are used to do LRA. This model, which we have called Logit, has been used to be informed about the performance of SVM.

**Model 3:** The Analysis with Feature Selection. In Model 3, LRA is used as the feature selection technique. Thanks to this analysis, the sub-set of features that will provide useful information was determined and SVM model was used. This model has been named Logit + Grid SVM.

### 4.3 Empirical Results and Discussion

Empirical results are analyzed under three main sections: The titles are: (1) *Logistic Regression Model Output*, (2) *SVM Models Output* (3) *The Performances of Models*.

#### 4.3.1 Logistic Regression Model Outputs

LRA takes the cumulative logistic function as the basis. This function, when the financial characteristics of the firms are given, gives the probability of whether the firm will be included in the distressed or non-distressed class. The empirical results of this model are presented in Table 3.

$x_1$  : asset growth,  $x_{19}$  : real operating profit margin,  $x_{17}$  : net profit margin,  $x_{21}$  : gross real operating profit margin,  $x_{23}$  : current ratio, and  $x_{22}$  : quick ratio in the model have been found to be significant at the 95% confidence level. The B value in the table indicates the coefficients of the logit model. The obtained logit model according to these results can be written as follows:

$$L_i = -1.031 - 1.761x_1 - 1.947x_{19} - 1.750x_{17} + 0.746x_{21} + 1.465x_{23} - 2.728x_{22}$$

It is seen that the prediction model is completely meaningful according to the statistical results ( $-2 \text{ Log Likelihood} = 86.949$ ;  $\chi - \text{Squared} = 12.493$ ; *degrees of freedom (d.f.)* = 8; *p value* = 0.131). From the statistical results of the coefficients ( $\chi - \text{Squared} 100,654$ ; *degrees of freedom (d.f.)* 6; *p value* = 0.000), it is concluded that the coefficients are significant. For the obtained Logit model, it is interpreted that the independent variables can explain 69.5% of the variability (*Nagelkerke R-Square* = 0.695) in the financial situations of the firms.

To calculate the probability of whether a firm is financially non-distressed, the relevant financial ratios of the firm are placed in the  $L_i$  function. The probability

**Table 3** Logistic model outputs

	$\beta$	Standard error	Wald	d. f	p-value	Exp ( $\beta$ )	95% confidence interval	
							Lower bound	Upper bound
$x_1$	-1.761	0.408	18.664	1	0.000	0.172	0.077	0.382
$x_{19}$	-1.947	0.767	6.453	1	0.011	0.143	0.032	0.641
$x_{17}$	-1.750	0.772	5.136	1	0.023	0.174	0.038	0.789
$x_{21}$	0.746	0.377	3.916	1	0.048	2.108	1.007	4.412
$x_{23}$	1.465	0.705	4.310	1	0.038	4.326	1.085	17.242
$x_{22}$	-2.728	1.096	6.197	1	0.013	0.065	0.008	0.560
Constant	-1.031	0.355	8.455	1	0.004	0.357	-	-

value corresponding to these numbers is calculated using  $P(L_i) = \frac{1}{1+e^{-L_i}}$  equation. When this value is higher than 0.5, it is decided that the firm will be non-distressed; otherwise, it will be distressed.

### 4.3.2 SVM Models Output

Under this title, the classification performances of Logit + Grid SVM models in which we have applied the LRA as the feature selection technique are compared. In addition to the optimization of SVM parameters, it is concluded that the optimal feature sub-set selection affects the classification success of SVM. The analysis outputs presented in Fig. 5 show that the parameters of SVM can affect the results. As has been mentioned in previous sections, when C and gamma values are set to be very high causes over-fitting error. In the analyses done on the test data, the classification success of the method decreases. When the constant value for C is determined to be  $2^5$  and when we look at the cross-validity rate that it takes for all values in the determined interval for gamma value, the cross-validation rate decreases to around 60%—the values shown on the blue line—at too high or too low values. This situation applies to both model propositions.

When we look at Fig. 5a, which shows Grid SVM results, the highest accuracy rate is 87.21%. It is seen that this accuracy rate is achieved when 2048 values for C and  $1.2207e-04$  values for gamma are assigned. In Fig. 5b, the impact of C and gamma on classification success in Logit + Grid SVM model is seen. Here, the highest cross-validation rate is 90.06% and this rate has been obtained at 256 value for C and 0.002 value for gamma. Another noteworthy point here is that the addition of feature selection stage to the analysis has increased the maximum value of cross-validation from 87.21 to 90.06%. In Table 4, a brief assessment of the effects of feature selection on SVM results was made. The values on the table are the values obtained by running both models 100 times on the test data set. As is indicated by

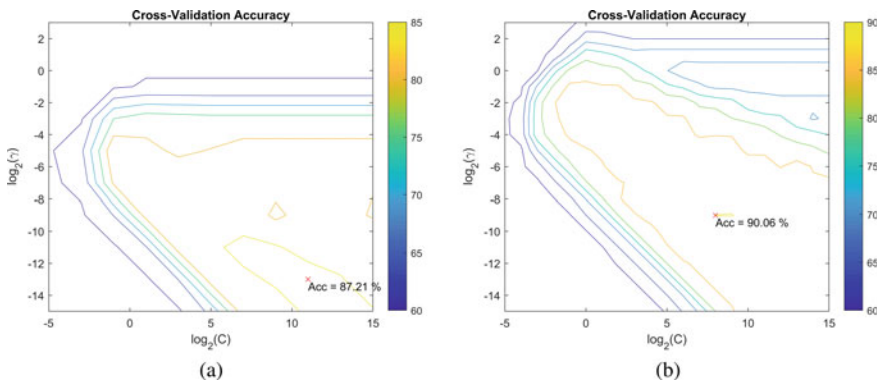


Fig. 5 Grid SVM (a) and Logit + Grid SVM Models Cross Validation Rates Graphs (b)

**Table 4** Empirical results regarding models

Models	Accuracy rate			Cross-validation rate		
	Mean	Standard Deviation	Max	Mean	Standard deviation	Max
Grid SVM	83.28%	(±0.0597)	90.63%	70.39%	(±11.25)	87.21%
Logit + Grid SVM	85.44%	(± 0.5520)	93.75%	74.80%	(±11.63)	90.12%

the results, the accuracy rate for SVM after feature selection increased from 83.28 to 85.44%. The cross-validation rate, which is a more reliable rate, increased from 70.39 to 74.80%.

### 4.3.3 Performances of the Proposed Models

It is seen that some different performance criteria are used in comparing the classification performances of the proposed models. Table 5 presents the results for the selected performance criteria. The accuracy rate of Logit + Grid SVM for training and the test sets are 94.24% and 93.75%, respectively. It can also be said that this model has a remarkably high sensitivity for both the training and the test set at a rate of 93.75% and 94.44%, respectively. The highest value of the specificity rate indicates the accuracy of the classifying model has been given by the logit model. It is the certainty rate which gives information about how many of the estimations of financial distress are real. The highest certainty value, too, has been obtained through Logit + Grid SVM. MCC value, which we have preferred for the situations in which the values in the confusion matrix are not distributed evenly, also provides information about the quality of the classifier. The highest MCC value again belongs to Logit + Grid SVM. It can be said that all three models are useful and produce classifiers with considerably high performances. As for the generalization capacity of the models, the relatively higher difference between the accuracy rates of the Logit model on the training data set and test data set indicates that its generalization performance is low.

**Table 5** Performance of the proposed models

	Grid SVM		Logit		Logit + Grid SVM	
	Training	Test	Training	Test	Training	Test
Accuracy	0.9282	0.9063	0.9000	0.8000	0.9424	0.9375
Sensitivity	0.9310	0.9000	0.8545	0.6153	0.9375	0.9444
Specificity	0.9268	0.9091	0.9294	0.9411	0.9452	0.9286
MCC	0.8539	0.7896	0.7893	0.6018	0.8831	0.8730
Precision	0.9000	0.8182	0.8867	0.8888	0.9615	0.9444

For precision, the Grid SVM yielded the lowest rate for the test set. This value is lower than the Logit has. Although it is shown in this study that logistic regression provides significant information with regard to the selection of the new feature sub-set, it is also seen that the performance of SVM operated by this new feature sub-set has increased.

## 5 Conclusion and Future Work

Since the financial distress of firms does not only affect the firm but also has an impact on the whole economy, financial distress prediction is a critically important subject, which has been frequently studied. In recent years, SVM has been commonly used in financial distress prediction studies. The financial distress model with SVM has been compared to other machine learning methods, it has been shown to yield good results. In the present study, it is aimed to make distress prediction by SVM. C and gamma parameters, which are considered as two significant parameters of SVM, are optimized by using grid search technique. It is shown to what extent the results are affected as a result of defining the relevant parameter pair correctly. Besides, it was seen that feature selection for SVM is another factor that significantly affects the results. To understand how feature selection affects classifying performance, the logistic regression analysis has been done. There are two reasons why this method has been chosen in the study: the first is that LRA does not require strict assumptions as in multiple-variable statistical techniques and it can be used as a feature selection technique; the second reason is that we wanted to compare the results of the logistic regression analysis to those of SVM.

Financial distress prediction is made based on a real data set of firms (172 firms) traded in the BIST share market between 2010 and 2017. The proposed models are compared based on this real data set. When the results of these proposed models are compared, it is concluded that SVM, which allows parameter optimization and feature selection, has a better success. As a consequence, a useful early warning model in financial distress prediction problem through SVM, a newly developed technique, is presented in the study.

## References

- Agostini M (2018) Corporate financial distress: going concern evaluation in both international and US contexts. Springer
- Akkaya GC, Demireli E, Yakut ÜH (2009) İşletmelerde Finansal Başarısızlık Tahminlemesi: Yapay Sinir Ağları Modeli ile IMKB Üzerine Bir Uygulama. Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi 10(2):187–216
- Aktaş R (1993) Endüstri İşletmeleri İçin Mali Başarısızlık Tahmini: Çok Boyutlu Model Uygulaması, T. İş Bankası Kültür Yayınları, Genel Yayın No 323, Ankara



- Alifiah MN (2014) Prediction of financial distress companies in the trading and services sector in Malaysia using macroeconomic variables. *Procedia Soc Behav Sci* 129:90–98
- Altman EI, Zhang L, Yen J (2007) Corporate financial distress diagnosis in China. New York University Salomon Center Working Paper
- Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Financ* 23(4):589–609
- Altman EI (1993) Corporate financial distress and bankruptcy. Wiley, New York
- Atiya AF (2001) Bankruptcy prediction for credit risk using neural networks: a survey and new results. *IEEE Trans Neural Netw* 12(4):929–935
- Beaver WH (1966) Financial ratios as predictors of failure. *J Account Res* 4:71–111
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2(2):121–167
- Caesarendra W, Widodo A, Yang BS (2010) Application of relevance vector machine and logistic regression for machine degradation assessment. *Mech Syst Signal Process* 24(4):1161–1171
- Chang CC, Lin CL (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST)* 2(3):1–27
- Chen MY (2011) Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Syst Appl* 38(9):11261–11272
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press
- Dambolena IG, Khoury SJ (1980) Ratio stability and corporate failure. *J Financ* 35(4):1017–1018
- Deakin EB (1972) A discriminant analysis of predictors of business failure. *J Account Res* 10(1):167–179
- Ding Y, Song X, Zeng Y (2008) Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Syst Appl* 34:3081–3089
- Doğan S, Koçak D, Atan M (2019) Support vector machines and logistic regression analysis on predicting financial distress model. In: International Conference on Data Science, Machine Learning and Statistics. pp 292–295
- Doğan S (2020) Optimal Parametre ve Özellik Seçimi ile Destek Vektör Makinesi Kullanılarak Finansal Başarısızlık Tahmini (Doktora Tezi), Gazi Üniversitesi, Ankara
- Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 35(5–6):352–359
- Etemadi H, Rostamy A, Dehkordi H (2009) A genetic programming model for bankruptcy prediction: empirical evidence from Iran. *Expert Syst Appl* 36:3199–3207
- Fan A, Palaniswami M (2000) Selecting bankruptcy predictors using a support vector machine approach. In: Proceeding of the International Joint Conference on Neural Network vol 6, pp 354–359
- Fletcher D, Goss E (1993) Forecasting with neural networks: an application using bankruptcy data. *Inf Manag* 24(3):159–167
- Gepp A, Kumar K, Bhattacharya S (2010) Business failure prediction using decision trees. *J Forecast* 29:536–555
- Gepp A, Kumar K (2008) The role of survival analysis in financial distress prediction. *Int Res J Financ Econ* 16:1450–2887
- Gunn SR (1998) Support vector machines for classification and regression. *ISIS Tech Rep* 14(1):5–16
- Hashi I (1997) The economics of bankruptcy, reorganization, and liquidation: lessons for East European Transition Economies. *Russ East Eur Financ Trade* 33(4):6–34
- Hsu CW, Chang CC, Li CJ (2003) A practical guide to support vector classification. Available from <http://www.csie.ntu.edu.tw/~cjlin/paper/guide/guide.pdf>
- Huang CL, Wang CJ (2006) A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst Appl* 31(2):231–240
- Jo H, Han I, Lee H (1997) Bankruptcy prediction using case-based reasoning, neural network and discriminant analysis for bankruptcy prediction. *Expert Syst Appl* 13(2):97–108

- Kumar P, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques. *A Rev Eur J Oper Res* 180:1–28
- Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 34(1):366–374
- Li H, Sun J, Wu J (2010) Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Syst Appl* 37(8):5895–5904
- Li H, Sun J (2012) Forecasting business failure: the use of nearest-neighbour support vectors and correcting imbalanced samples: evidence from the chinese hotel industry. *Tour Manage* 33(3):622–634
- Liang D, Lu CC, Tsai CF, Shih GA (2016) Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study. *Eur J Oper Res* 252:561–572
- Liang L, Wu D (2005) An application of pattern recognition on scoring chinese corporations financial conditions based on backpropagation neural network. *Comput Oper Res* 32(5):1115–1129
- Liao Y, Fang SC, Nuttle HLW (2004) A neural network model with bounded-weights for pattern classification. *Comput Oper Res* 31:1411–1426
- Lin SW, Lee ZJ, Chen SC, Tseng TY (2008) Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl Soft Comput* 8:1505–1512
- Lin WY, Hu YH, Tsai CF (2012) Machine learning in financial crisis prediction: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev* 42(4):421–436
- Lincoln M (1984) An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk. *J Bank Finance* 8(2):321–340
- Liu H, Motoda H (1998) Feature extraction, construction and selection: a data mining perspective. Springer Science & Business Media
- Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst, Man, Cybern Part B (Cybern)* 39(2):539–550
- Meyer AD (1982) Adapting to environmental jolts. *Adm Sci Q* 515–537
- Meyer PA, Pifer H (1970) Prediction of bank failures. *J Financ* 25:853–868
- Min JH, Lee YC (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Syst Appl* 28(4):603–614
- Min SH, Lee J, Han I (2006) Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Syst Appl* 31:652–660
- Odom M, Sharda R (1990) “A neural networks model for bankruptcy prediction”. In: *Proceedings of The IEEE International Conference on Neural Network* 2:163–168
- Ohlson J (1980) Financial ratios and the probabilistic prediction of bankruptcy. *J Account Res* 18(1):109–131
- Özdemir FS (2011) Finansal Başarısızlık ve Finansal Tablolara Dayalı Tahmin Yöntemleri. *Ank: Siyasal Kitapevi* 82(33–37):106–108
- Pardo M, Sberveglieri G (2005) Classification of electronic nose data with support vector machines. *SensS Actuators* 107:730–737
- Park C-S, Han I (2002) A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Syst Appl* 23:255–264
- Pindado J, Rodrigues L (2005) Determinants of financial distress costs. *Fin Markets Portfolio Mgmt* 19(4):343–359
- Piramuthu S (2004) Evaluating feature selection methods for learning in data mining application. *Eur J Oper Res* 156:483–494
- Saeyns Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517
- Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min Knowl Discov* 1:317–327
- Shin KS, Lee TS, Kim HJ (2005) An applications support vector machines in bankruptcy prediction model. *Expert Syst Appl* 28:127–135

- Shin KS, Lee YJ (2002) A genetic algorithm application in bankruptcy prediction modeling. *Expert Syst Appl* 23:321–328
- Smola A, Schölkopf B (1997) On kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica* 22:211–231
- Srang G (1986) Introduction to applied mathematics. Wellesley-Cambridge Press, Wellesley, MA
- Sun J, Li H, Huang QH, He KY (2014a) Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowl Based Syst* 57:41–56
- Sun J, Li H (2011) Dynamic financial distress prediction using instance selection for the disposal of concept drift. *Expert Syst Appl* 38:2566–2576
- Sun J, Li H, Huang QH, He KY (2014b) Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowl Based Syst* 57:41–56
- Tam K, Kiang M (1992) Managerial applications of neural networks: the case of bank failure predictions. *Manage Sci* 38(7):926–947
- Tam K (1991) Neural network models and the prediction of bank bankruptcy. *Omega* 19(5):429–445
- Tsai CF (2008) Financial decision support using neural networks and support vector machines. *Expert Syst J Knowl Eng* 25(4):380–393
- Vapnik VN (1995) The nature of statistical learning theory. Springer-Verlag, New York
- Woods K, Bowyer KW (1997) Generating ROC curves for artificial neural networks. *IEEE Trans Med Imaging* 16(3):329–337
- Wu CH, Tzeng GH, Goo YJ, Fang WC (2007) A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Syst Appl* 32(2):397–408
- Wu W, Cheng V, Lee S, Tan TY (2006) Data preprocessing and data parsimony in corporate failure forecast models: evidence from Australian materials industry. *Account Financ* 46:327–345
- Yilmaz I (2009) Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat Landslides (Tokat-Turkey). *Comput Geosci* 35(6):1125–1138
- Zhang G, Hu MY, Patuwo BE, Indro DC (1999) Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *Eur J Oper Reseach* 116:16–32
- Zhongsheng H, Yu W, Xiaoyan X, Bin Z, Liang L (2007) Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Syst Appl* 33(2):434–440
- Zhou L, Lu D, Fujita H (2015) The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowl-Based Syst* 85:52–61
- Zmijewski ME (1984) Methodological issues related to the estimation of financial distress prediction models. *J Account Res* 22:59–82

# Predicting Stock Returns: ARMAX versus Machine Learning



Darya Lapitskaya, Hakan Eratalay, and Rajesh Sharma

**Abstract** In the modern world, online social and news media significantly impact society, economy and financial markets. In this chapter, we compared the predictive performance of financial econometrics and machine learning and deep learning methods for the returns of the stocks of the S&P 100 index. The analysis is enriched by using COVID-19-related news sentiments data collected for a period of 10 months. We analysed the performance of each model and found the best algorithm for such types of predictions. For the sample we analysed, our results indicate that the autoregressive–moving-average model with exogenous variables (ARMAX) has a comparable predictive performance to the machine and deep learning models, only outperformed by the extreme gradient boosted trees (XGBoost) approach. This result holds both in the training and testing datasets.

**Keywords** Sentiment analysis · Machine learning · ARMAX · Stock returns prediction · Deep learning · COVID-19

## 1 Introduction

In the modern world, social and news media significantly impact society and the economy (Bruhn et al. 2012, Hanusch and Tandoc 2019). They change the companies' business models and affect their performance and reputation as the opinions about a product or service now can be freely shared online. Hence, media also affect the stock

---

D. Lapitskaya · H. Eratalay (✉)  
School of Economics and Business Administration, University of Tartu, Tartu, Estonia  
e-mail: [hakan.eratalay@ut.ee](mailto:hakan.eratalay@ut.ee)

D. Lapitskaya  
e-mail: [darya.lapitskaya@ut.ee](mailto:darya.lapitskaya@ut.ee)

R. Sharma  
Institute of Computer Science, University of Tartu, Tartu, Estonia  
e-mail: [rajesh.sharma@ut.ee](mailto:rajesh.sharma@ut.ee)

markets and stock prices. Many researchers found dependencies between information and media and the company's performance at the stock market (Steyn et al. 2020, Khan et al. 2020, Coyne et al. 2017).

The impact of information on the stock market and stock price volatilities has been investigated for many years (Malkiel et al. 2003). According to the earlier research and efficient market hypothesis, stock market prices are more affected by new information than by the present and past prices (Arafat et al. 2013). Later on, it was proved that the public mood measured from posts on social media was correlated with the market prediction (Arafat et al. 2013). Also, it was demonstrated that news and social media sentiments could predict future stock returns (Leung et al. 2019). According to the research in Mohan et al. (2019), there is a strong correlation between the stock price volatilities and the news articles. Hence, it can be concluded that online social media sentiments could be used to predict stock returns.

The prediction of stock returns is usually performed with different econometrical models, such as the autoregressive–moving-average with exogenous inputs (ARMAX). However, nowadays, machine learning (ML) (Shah et al. 2018), deep learning (Abe and Nakayama 2018), and graph neural networks (Sharma and Sharma 2020) are actively used in financial econometrics and forecasts. The usage of ML forecasts brings significant financial benefits to investors, as in some cases, such methods have doubled the predictive performance of leading regression-based methodologies (Gu et al. 2018). The interest in comparing ARMAX, ML and deep learning methods' predictive performance rests in the possible disadvantages of these methods. On the one hand, ARMAX models are restricted with the stationarity and invertability conditions on the coefficient estimators, which would trade off the predictive power with the stationary behaviour of the predicted variable. On the other hand, the ML and deep learning methods focus on better predictive performance considering many lags and/or functional forms of the variables but leaving the interpretability of the coefficients in the back plan. So, when the interest is in predicting the future behaviour of returns, it is expected that ML and deep learning methods would outperform the ARMAX method in terms of predictive performance.

This chapter gives an example of the mentioned comparison using highly volatile data due to the effect of COVID-19. In particular, this study is dedicated to analysing the impact of COVID-19-related news on the Standard and Poor's 100 companies (S&P 100) stock returns by collecting news article for the period of 10 months. Specifically, we analyse stock returns predictions with sentiments scores by comparing two different prediction methods. The first one belongs to the traditional econometric domain (ARMAX modelling), and the second one is from the domain of machine learning (KNN, XGBoost and a deep learning neural network (LSTM)).

The comparison between different prediction techniques showed that the XGBoost had better predictive performance than the ARMAX model, while KNN and LSTM performed worse. Although the result is interesting that the ARMAX model performed well despite the restrictions on the parameters than KNN and LSTM, we acknowledge that the small sample size might be the culprit. The findings of this research contribute to the literature and allow understanding the impact

of COVID-19-related news articles on the stock markets. Moreover, the application of both sentiment analysis and ML prediction techniques helps to create a more precise returns prediction model.

The remainder of the chapter is organized in the following way: in the next section, the theoretical aspects of stock return predictions are presented as well as the analysis of previous research done regarding the sentiments impacts on the stock market. The third section is dedicated to the research method and its' methodology. In the fourth section, the results are presented together with a discussion followed by the study conclusions and future research perspectives in the last section.

## 2 Literature Review

Stock prices forecasting is one of the critical fields in the financial econometrics, hence the stock price and returns predictions are essential parts of the stock market analysis (Kordonis et al. 2016). Usually, the forecast is done with the historical stock price data, however media sentiments also affect the stock prices fluctuations and could be included in the prediction models (Shah et al. 2018). In general, sentiment scores are calculated based on news articles or social media/microblogging posts. The forecasting with microblogging data was studied by different researchers. For example, it was shown in Kordonis et al. (2016) that there was a correlation between Twitter sentiments and stock prices. Also, the research Cazzoli et al. (2016) demonstrated that Twitter posts related to corporations could predict the financial market.

Moreover, stock-specific sentiments have a bigger impact on returns than market-specific sentiments (Anusakumar et al. 2017). Furthermore, it was shown in Wolf and O. Bergdorf (2019) that the sentiments derived from Twitter were useful in the individual stock returns predictions. In a recent study done during the COVID period, it is shown that tweets containing the term *stocks* have a substantial decline in log returns for US indices (Goel et al. 2020). Also, there is a significant correlation between the changes in stock prices and the publication of news articles (Mohan et al. 2019). Using news sentiments as a predictor variable leads to a directional accuracy of 70.59 percent in short-term stock price movement trends prediction (Shah et al. 2018).

Prediction computations can be done with econometrical and statistical approaches or with the adaptation of ML and deep learning algorithms. And nowadays, more and more researchers prefer using the ML algorithm for predictions or combine them with traditional methods. The usage of ML algorithms allows improving accuracy and overcome limitations of common econometrical models (Rossi 2018). Moreover, the ML algorithms outperform the benchmark buy-and-hold strategy in the real-life simulations and the gradient boosting machine performs the best from the perspective of the statistical and economic evaluation criteria (Nevasalmi (2020)).

Deep Learning algorithms also show potential in stock returns predictions as they can analyse complex patterns and interactions in the dataset (Vargas et al. 2017).

Also, deep neural networks could outperform shallow neural networks, and some of them could even outperform representative machine learning models (Abe and Nakayama 2018). Typically, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) can be used for stock market forecast (Vargas et al. 2017). The above-mentioned deep and ML algorithms could be combined with sentiment analysis to receive better accuracy of returns. For example, the RNN models that used news articles text in the input performed better than ones that predicted future stock prices based only on historical stock prices Mohan et al. (2019). CNN outperform RNN on catching textual semantics, and RNN better catches the context information and performs better in modelling complex temporal characteristics (Vargas et al. 2017).

### 3 Methodology

This chapter focuses on predictive modelling of Shmueli (2011) and predictive powers of each chosen stock return prediction method.

However, before constructing the models, it is crucial to understand the timing of the variables. The return vector is calculated using the log-difference formula, which is commonly used in finance:

$$r_t = \log(P_t/P_{t-1}) = \log(P_t) - \log(P_{t-1}) \quad (1)$$

where  $P_t$  is the adjusted closing price of the stock market index. Therefore, a return is calculated as the change between the closing values of subsequent trading days. Consequently, any news published in the news sources played a role in deciding what would be the next closing price; hence, the return.

The model should also consider that returns are available only for the trading days, but the news is published daily, including the weekends and holidays. Ignoring the news data from the non-trading days would result in losing important information. The news that appears during the weekends and holidays affect the investors' behaviour, and this is reflected in the stock prices in the first subsequent trading day. The return that is calculated as the log-difference of closing prices before and after a weekend or holiday contains the news information during these days. That's why, in this study, the news data from weekends and holidays is merged to the first following trading day's news data. In fact, these effects are called weekend (a.k.a. Monday) and holiday effects in financial econometrics literature (Basher and Sadorsky 2006, Marrett and Worthington 2009). Hence, a proper model should also consider the accumulated impact of the news from weekends and holidays.

Another issue to focus on is the asymmetric effect of the news on returns. It has been documented that the negative return shocks affect the returns and volatilities differently than the positive return shocks (see Maheu and McCurdy 2004, Puzanova and Eratalay 2021, Ensor et al. 2020, Engle and Ng 1993). The models in this paper

incorporate this asymmetric effect of the news by considering positive and negative news sentiment scores separately.

### 3.1 ARMAX Model

Taking into account the discussion above and adapting the approach of Puzanova and Eratalay (2021), the following ARMAX model is constructed:

$$\begin{aligned}
 r_t = & \mu + \mu_W D_t^W + \mu_{NO} D_t^{NO} + \mu_{PO} D_t^{PO} \\
 & + \sum_{i=1}^p \beta_i r_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \\
 & + \delta_{1N} News_t^N + \delta_{1P} News_t^P \\
 & + \delta_{2N} News_{t-1}^N + \delta_{2P} News_{t-1}^P \\
 & + \gamma_{1N} Newscount_t^N + \gamma_{1P} Newscount_t^P \\
 & + \gamma_{2N} Newscount_{t-1}^N + \gamma_{2P} Newscount_{t-1}^P
 \end{aligned} \tag{2}$$

where  $D_t^W$  is a dummy variable for the weekend and holiday effects;  $D_t^{NO}$  and  $D_t^{PO}$  are dummy variables for negative and positive outliers<sup>1</sup>, respectively;  $News_t^N$  and  $News_t^P$  are negative and positive merged news sentiment scores, respectively; and  $Newscount_t^N$  and  $Newscount_t^P$  are the number of negative and positive news that occurred at day  $t$ . The ARMAX orders  $p$  and  $q$  are chosen by comparing the AIC values of the model estimates. The autoregressive parameters  $\beta_i$  and moving-average parameters  $\theta_j$  are restricted to satisfy the stationarity and invertability restrictions, respectively.

### 3.2 Sentiment Analysis

To improve the prediction accuracy, we use news articles’ sentiment score as prediction variables. Hence, each article of the dataset was analysed and assessed with a sentiment analysis algorithm. In general, the sentiment analysis is used to identify opinions expressed in the textual form, and it is based on a natural language processing algorithm where each word of the text has its sentiment score (positive, neutral or negative) (Luo et al. 2013). Sentiment analysis can be performed using supervised and non-supervised approaches. The non-supervised approach represents the classification done ‘based on a dictionary-based approach to convert the qualitative news articles into a quantitative measure’ (Li et al. 2018). The supervised

---

<sup>1</sup> Outliers were identified using Hampel filtering and Python calculations.



approach of sentiment analysis uses historical trends and news patterns and creates training data by automatic labelling of news and social media posts (Yadav and A. Kumar 2019). In this research, we applied Valence Aware Dictionary and Sentiment Reasoner (VADER) method. This is a semi-supervised algorithm, which is a simple rule-based model for general sentiment analysis (Hutto and Gilbert 2015) to the collected news articles. For each day, we calculated average news score. All the news articles with neutral sentiment scores (score equals to zero) were dropped from the calculation. Also, we have counted numbers of positive and negative news per day to use them as separate independent variables.

### 3.3 Machine Learning and Deep Learning Modelling

As an alternative to ARMAX model in this study, we used two ML algorithms and one deep learning neural network to predict stock returns: eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbours (KNN) and Long Short-Term Memory (LSTM) regression models. These algorithms were chosen due to their high accuracy in the regression forecasting. For these algorithms, we used the same dataset, model and dataset split ratio as in ARMAX model to create fair conditions for comparison. All the computations were performed in *Python*.

**XGBoost:** The first algorithm we applied to the chosen regression model was XGBoost ML algorithm designed for efficacy, computational speed and model performance that demonstrates good performance in solving regression and classification problems (Malik et al. 2020). XGBoost is a tree boosting method that is considered a highly effective and widely used ML approach that can solve practical problems using a minimal amount of resources (Chen and Guestrin 2016). While building the regression XGBoost uses a loss function to evaluate the prediction model. In particular, the XGBoost prediction model was constructed by using the *xgboost* library and the *xgboost.XGBRegressor* function.

**KNN:** The second ML algorithm we used for returns prediction was KNN. KNN algorithm is a widespread ML algorithm for regression analysis. Its' choice is justified by its simplicity and easy adaptation process, hence it is commonly used for time series analysis and forecast (Ban et al. 2013). In the KNN regression algorithm, the dependent variable of a time series forecast is described as a sequence of interval scaled values. Then, based on the pattern, the KNN algorithm identifies similar past patterns and combines their future values to form predictions (Ban et al. 2013). The KNN model was created with *sklearn.neighbors* library and *KNeighborsRegressor* function.

**LSTM:** Finally, we created the LSTM regression model. LSTM is a type of recurrent neural network (one of the general classes of neural networks) deep learning-based algorithm which is commonly used in times series forecasting (Elsworth and S. Güttel 2020, Sherstinsky 2020). The LSTM-based regression algorithm is a multi-step

univariate forecast algorithm that demonstrates a good accuracy in processing the dependency among the dependent variables (Siami Namini et al. 2018). The LSTM-based regression was estimated with *Keras* library.

The main issue with ML and deep learning algorithms is that they usually require big volumes of data to properly learn and provide accurate results. In this study, we were using a relatively small dataset, so we were also testing whether the chosen algorithms could outperform ARMAX-based prediction with the small volumes of information to process.

### 3.4 Dataset

The dataset used for ARMAX and ML modelling consists of S&P 100 historical data, the negative and positive sentiment scores, and the number of news.

The news data was collected using the web scraping method. In total, we have collected over 6000 news articles related to the COVID-19 pandemic that was later cleaned, pre-processed (duplicates were deleted, and the dataset was sorted by date) and were used to conduct sentiment analysis. This dataset covers the period from 27.01.2020 to 10.12.2020. Even though the spread of COVID-19 started in December 2019, there were not enough news articles data for that time to make a sufficient analysis. Hence, the data collection started from 27.01.2020 to have the sufficient number of news regarding COVID-19 published on a daily basis. The dataset gives 223 observations for the returns and 11 exogenous variables. The return observations contain outliers identified by Hampel filtering that are shown in Fig. 1a. To have a visual presentation of the returns distribution, a quantile-to-quantile (QQ) plot is presented in Fig. 1b.

The descriptive statistics can be found in Table 1. The skewness and, in particular, the kurtosis values suggest that the returns are not normally distributed. This is

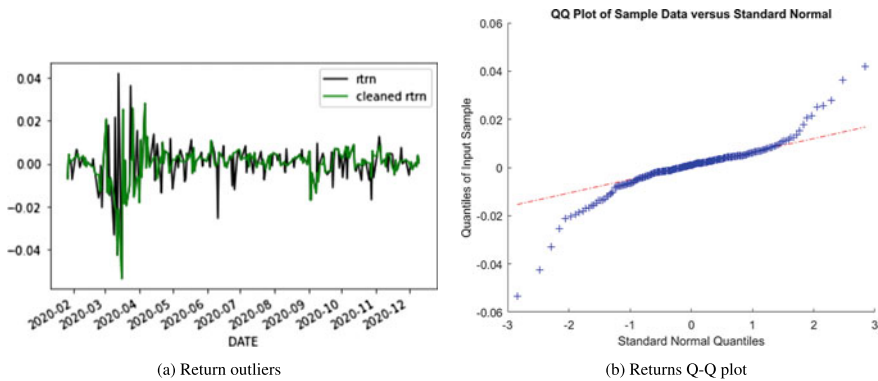


Fig. 1 Return outliers and QQ plot of the returns

**Table 1** Descriptive statistics of the SP100 returns for the data period

Mean	Median	Variance	Skewness	Kurtosis	JB test pval.
0.0002	0.0010	0.0001	-0.6850	9.9869	0.0010

confirmed by the p-value of the Jarque–Bera test of normality for the returns. What is partially responsible for the high kurtosis is the existence of outliers. In the QQ plot presented in Fig. 1b, it can be seen that there are some positive and negative outliers at the top right and bottom left of the figure, respectively. These outliers are identified by the Hampel filter as shown in Fig. 1a, where time series plots of the true returns and the returns cleaned from these outliers are presented. It should be noted that the identified outliers in the return series are not removed or smoothed out. Instead, as shown in Sect. 3.1, we add dummy variables to the model to control for the positive and negative outliers.

## 4 Results and Discussions

We choose the mean absolute error (MAE) and root mean squared error (RMSE) as comparison indicators of predictive performances of the models. The dataset was split in training and testing dataset to perform the prediction modelling. The training dataset includes the period from 27.01.20 to 31.08.20 (68 percent of the whole dataset) and the test dataset covers the period from 01.09.2020 to 10.12.20 (32 percent of the whole dataset). This corresponds to almost 70:30 percent split, which is a splitting ratio usually used in ML.

The ARMAX model orders of Eq. 2 were chosen as  $p=1$  and  $q=1$ , by identifying the ARMAX specification that gave the lowest AIC value. In Fig. 2, the histogram of the residuals, in modulus, of the ARMAX model estimation is plotted. Most of the residuals are concentrated towards zero, while a few of the residuals lie in the tail of the histogram. A Ljung–Box test on the residuals up to 15 lags showed that the autocorrelation in the returns is successfully captured by the ARMAX model.

Despite the fact that the ARMAX model requires stationarity and invertibility restrictions on the model parameters (see Sect. 3.1), it is interesting to see that it learned on the training set well. Figure 3a shows that the ARMAX model was able to predict closely most of the highest and lowest returns, probably because of the dummy variables for the outliers in the model. However, its predictions were smoother than the true returns and couldn't catch the variation in it so well. In contrast, the ARMAX model produced better results for the test dataset. As Fig. 3b suggests, it wasn't able to predict the high and low points as much in the test dataset, but the spread of the predictions was slightly higher, which resulted in better predictive performance. The prediction MAE results were *0.00511* for the training dataset and *0.00487* for the test

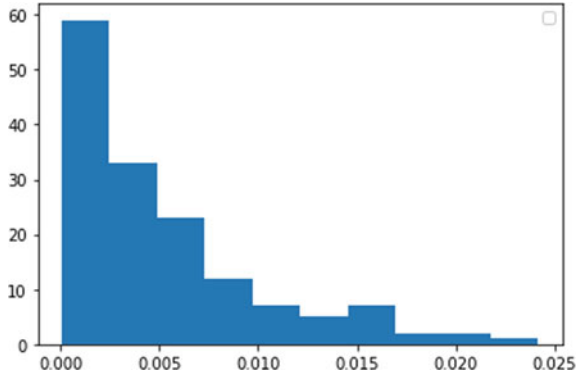


Fig. 2 Returns residuals

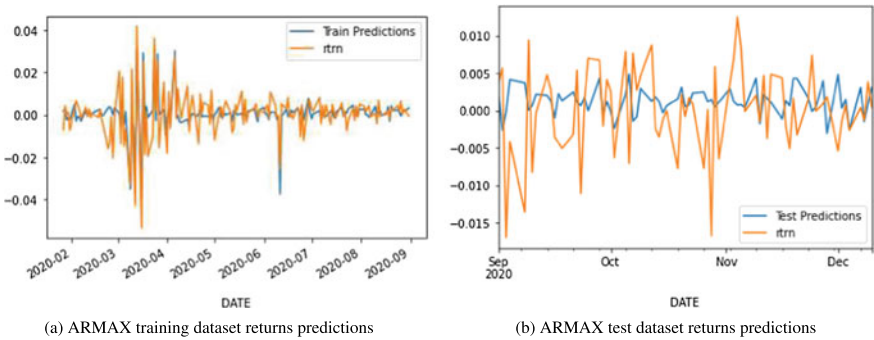


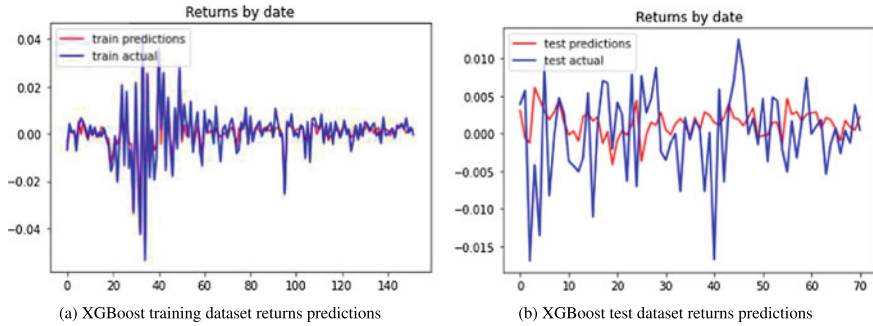
Fig. 3 ARMAX prediction results

Table 2 Results for stock returns prediction models

Method	MAE training	MAE test	RMSE training	RMSE test
ARMAX	0.00511	0.00487	0.00718	0.00640
XGBoost	<b>0.00316</b>	<b>0.00479</b>	<b>0.00433</b>	<b>0.00639</b>
KNN	0.00544	0.00534	0.00893	0.00682
LSTM	0.01049	0.00742	0.01347	0.00854

dataset. Meanwhile, not all the ML algorithms outperformed the ARMAX model in predictive performance (see Table 1).

The results in Table 2 show that the XGBoost algorithm gave the best prediction result on the given dataset. Figure 4a and b shows the XGBoost predictions of the returns and the true returns in the training set and testing set, respectively. The XGBoost algorithm was able to learn very well from the training set, although the predictive performance in the training set is not so much above the one for the ARMAX model.



**Fig. 4** XGBoost prediction results

The deep learning (LSTM) model could not properly learn due to the small sample size, and it was outperformed by ARMAX predictions both in the training and testing sets. The results were worse for KNN: the MAE and RMSE results for the KNN were almost double the ones for the ARMAX predictions. The problem could again be connected to the small dataset size.

To summarize, we can see that the XGBoost algorithm outperformed ARMAX for the training dataset, but gave a similar performance with ARMAX in the testing dataset. The other ML approaches couldn't perform that well. One could conclude here that the XGBoost was the most suitable algorithm for this specific sample, followed by the ARMAX model. It is important to point out here that the findings in this analysis only apply to this particular data, which is not a very large sample, volatile and with some outliers.

## 5 Conclusions and Future Research

The impact of information on stock markets was investigated by many researchers. The previous studies suggest that there is a correlation between news and media sentiments and stock returns. This chapter contributes to the literature in several dimensions. On the one hand, the effect of the news sentiments on the returns of the SP100 index was analysed considering the possibility of the asymmetric effect of negative and positive news. On the other hand, the analysis was conducted using the period when the markets were very volatile and very sensitive to the news about COVID-19. Lastly, the analysis compared the predictive performance of ARMAX and the ML algorithms. The results of the analysis demonstrate that the sentiment score inclusion and usage of the ML algorithm significantly increase the accuracy of the prediction. We found that the XGBoost prediction model showed the best results and had the highest predictive power. In terms of comparing the predictive performances of the mentioned models, this is not an exhaustive study. Therefore, the findings only relate to the specific data and period under consideration. Future

research could extend the comparison of the predictive performances by increasing the sample size of the returns and considering many different data characteristics related to the distribution of the returns. Moreover, an exhaustive simulation study on the comparison of these methods using data with many different statistical properties is planned by the authors in future research. There could be many factors that could be considered for this comparison, some of which are the volatility of the data at hand, the number of outliers in the training and testing sets, the autocorrelation structure, structural breaks and misspecification of the distribution.

**Acknowledgements** This work has been supported by GrowInPro (Horizon 2020), SoBigData++ (Horizon 2020) and SAI (CHIST-ERA) Projects.

## References

- M. Abe and H. Nakayama. Deep learning for forecasting stock returns in the cross-section. 01 2018
- S. Anusakumar, R. Ali, and C. Hooy. The effect of investor sentiment on stock returns: Insight from emerging asian markets. *Asian Academy of Management Journal of Accounting and Finance*, 13:159–178, 01 2017
- J. Arafat, M. A. Habib, and R. Hossain. Analyzing public emotion and predicting stock market using social media. pages 265–275, 01 2013
- T. Ban, R. Zhang, S. Pang, A. Sarrafzadeh, and D. Inoue. Referential knn regression for financial time series forecasting. pages 601–608, 11 2013
- Basher SA, Sadorsky P (2006) Day-of-the-week effects in emerging stock markets. *Applied Economics Letters* 13(10):621–628
- M. Bruhn, V. Schoenmueller, and D. Schäfer. Are social media replacing traditional media in terms of brand equity creation? *Management Research Review*, 35:770–790, 08 2012
- L. Cazzoli, R. Sharma, M. Treccani, and F. Lillo. A large scale study to understand the relation between twitter and financial market. In *2016 third European network intelligence conference (ENIC)*, pages 98–105. IEEE, 2016
- T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016
- S. Coyne, P. Madiraju, and J. Coelho. Forecasting stock prices using social media analysis. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 1031–1038, 2017
- S. Elsworth and S. Güttel. Time series forecasting using lstm networks: A symbolic approach. 03 2020
- Engle RF, Ng VK (1993) Measuring and testing the impact of news on volatility. *The journal of finance* 48(5):1749–1778
- Ensor KB, Han Y, Ost diek B, Turnbull SM (2020) Dynamic jump intensities and news arrival in oil futures markets. *Journal of Asset Management* 21(4):292–325
- R. Goel, L. J. Ford, M. Obrizan, and R. Sharma. Covid-19 and the stock market: evidence from twitter. *arXiv preprint <http://arxiv.org/abs/2011.08717>*, 2020
- Gu S, Kelly B, Xiu D (2018) Empirical asset pricing via machine learning. *Globalization eJournal, International Political Economy*
- Hanusch F, Tandoc EC (2019) Comments, analytics, and social media: The impact of audience feedback on journalists’ market orientation. *Journalism* 20:695–713
- C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015

- W. Khan, M. a. Ghazanfar, M. A. Azam, A. Karami, K. Alyoubi, and A. Alfakeeh. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 03 2020
- J. Kordonis, S. Symeonidis, and A. Arampatzis. Stock price forecasting via sentiment analysis on twitter. 11 2016
- W. S. Leung, W. K. Wong, and G. Wong. Social-media sentiment, limited attention, and stock returns. *Economics of Networks eJournal*, 2019
- M. Li, C. Yang, J. Zhang, D. Puthal, Y. Luo, and J. Li. Stock market analysis using social networks. ACSW '18, New York, NY, USA, 2018. Association for Computing Machinery
- T. Luo, S. Chen, G. Xu, and J. Zhou. *Sentiment Analysis*, pages 53–68. 06 2013
- Maheu JM, McCurdy TH (2004) News arrival, jump dynamics, and volatility components for individual stock returns. *The Journal of Finance* 59(2):755–793
- S. Malik, R. Harode, and A. Kunwar. Xgboost: A deep dive into boosting (introduction documentation). 02 2020
- B. Malkiel. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17:59–82, 02 2003
- Marrett GJ, Worthington AC (2009) An empirical note on the holiday effect in the Australian stock market, 1996–2006. *Applied Economics Letters* 16(17):1769–1772
- S. Mohan, S. Mullanpudi, S. Sammeta, P. Vijayvergia, and D. Anastasiu. Stock price prediction using news sentiment analysis. pages 205–208, 04 2019
- Nevasalmi L (2020) Forecasting multinomial stock returns using machine learning methods. *The Journal of Finance and Data Science* 6:86–106
- Puzanova Y, Eratalay MH (2021) Effect of real estate news sentiment on the stock returns of swedbank and seb bank. *Econ Res Finan* 6(2):77–117 (2021). <https://doi.org/10.2478/erfin-2021-0005>
- A. G. Rossi. Predicting stock market returns with machine learning. 2018
- D. Shah, H. Isah, and F. Zulkernine. Predicting the effects of news sentiments on the stock market. 12 2018
- S. Sharma and R. Sharma. Forecasting transactional amount in bitcoin network using temporal gnn approach. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020
- A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 03 2020
- Shmueli G (2011) To explain or to predict? *Statistical Science* 25:01
- S. Siami Namini, N. Tavakoli, and A. Siami Namin. A comparison of arima and lstm in forecasting time series. pages 1394–1401, 12 2018
- D. H. W. Steyn, T. Greyling, S. Rossouw, and J. M. Mwamba. Sentiment, emotions and stock market predictability in developed and emerging markets. 2020
- M. Vargas, B. Lima, and A. Evsukoff. Deep learning for stock market prediction from financial news articles. pages 60–65, 06 2017
- F. Wolf and O. Bergdorf. Twitter sentiment and stock returns. 2019
- Yadav R, Kumar A (2019) News-based supervised sentiment analysis for prediction of futures buying behaviour. *IIMB Management Review* 31:04

# Analysing the Residential Market Using Self-Organizing Map



Olgun Aydin and Krystian Zieliński

**Abstract** Although the residential property market has strong connections with various sectors, such as construction, logistics, and investment, it works through different dynamics than other markets; thus, it can be analysed from various perspectives. Researchers and investors are mostly interested in price trends, the impact of external factors on residential property prices, and price prediction. When analysing price trends, it is beneficial to consider multidimensional data that contain attributes of residential properties, such as number of rooms, number of bathrooms, floor number, total floors, and size, as well as proximity to public transport, shops, and banks. Knowing a neighbourhood's key aspects and properties could help investors, real estate development companies, and people looking to buy or rent properties to investigate similar neighbourhoods that may have unusual price trends. In this study, the self-organizing map method was applied to residential property listings in the Trójmiasto Area of Poland, where the residential market has recently been quite active. The study aims to group together neighbourhoods and subregions to find similarities between them in terms of price trends and stock. Moreover, this study presents relationships between attributes of residential properties.

**Keywords** Self-organizing map · Real estate market · Clustering · Residential property prices

## 1 Introduction

In the broadest terms, residential property can be defined as a set of physical spaces that meet the accommodation needs of people. Residential property can also be considered an investment tool or a property that can be used as collateral when

---

O. Aydin (✉)  
Gdańsk University of Technology, Gdansk, Poland  
e-mail: [olgun.aydin@pg.edu.pl](mailto:olgun.aydin@pg.edu.pl)

K. Zieliński  
PwC Polska, Gdańsk, Poland  
e-mail: [krystian.zielinski@pwc.com](mailto:krystian.zielinski@pwc.com)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,  
[https://doi.org/10.1007/978-3-030-85254-2\\_28](https://doi.org/10.1007/978-3-030-85254-2_28)

465



necessary. The market for residential properties is based on the supply–demand balance (Aydin & Hayat, 2018).

Demand for residential properties is constantly increasing due to rapidly increasing urbanization and migration, and increasing demand significantly impacts on property prices. Prices can be obtained from various sources, such as real estate development companies, real estate agencies, newspapers, listing portals, real estate appraisal companies, notary records, and government institutions (Hepsen, 2015). Regarding residential property as an investment tool, investors aim to profit from the purchase–sale difference or to obtain a regular rental income from the property in which they invest (Erpolat Tasabat, 2018).

This study's primary purpose is to analyse the residential property market in Trójmiasto, a metropolitan area in the north of Poland consisting of the three cities Gdansk, Sopot, and Gdynia. Due to migration, increasing employment opportunities, foreign investment, and increasing life quality, Trójmiasto's residential market has grown dynamically in the last few years. One aim of this study is to analyse the secondary and primary markets separately from the perspective of sales prices. For this purpose, residential property listings were obtained from *trojmiasto.pl*, which is the biggest web portal in Trójmiasto. The dataset contains active listings from 2020.

The structure of the paper is as follows. The following (second) section provides an overview of the residential market in Trójmiasto, and the third section presents a literature review. Section 4 describes self-organizing maps and Sect. 5 explains the methodology used in the study and presents the findings. The final section concludes the study and recommends potential research.

## 2 Residential Market in Trójmiasto

Trójmiasto (Tricity, or Tri-City in English) is a metropolitan area in the north of Poland consisting of three cities, Gdańsk, Gdynia, and Sopot, as well as small towns in the surrounding area. Tri-City is located on the coast of Gdańsk Bay, by the Baltic Sea, in East Pomerania. According to official records, the population of Tri-City metropolitan area is over one million. The name Tri-City was used informally until 28 March 2007, when the Tri-City Charter (Karta Trójmiasto in Polish) was signed as a declaration of cooperation between cities in the region (Karta Trójmiasta, 2007).

Tri-City differs from the rest of Poland in its socio-economic aspects. Data obtained from the Polish Central Statistical Office show that in 2019, Poland's average gross income was 5,182 PLN, whereas in Gdańsk it was 6,154 PLN, in Gdynia 5,624 PLN, and in Sopot 6,064 PLN. Regarding unemployment statistics, Tri-City is quite different from the rest of Poland. The number of registered unemployed person per single job offer is 11.4 for all Poland, while for Gdańsk it is 7.25, for Gdynia 7.17, and for Sopot 8.36; this is connected to the rapid development of office space and the increasing number of international companies in Tri-City. Another factor making Tri-City wealthy is the touristic potential of the area, which is connected with the Baltic Sea. Especially Sopot is highly regarded as the summer capital of Poland.

**Table 1** Housing in 2019

Area	Number of flats per 1,000 inhabitants	Average flat size (square metre)	Number of flats built in last 3 years per 1,000 inhabitants
Poland	5.4	88.6	5.0
Gdańsk	15.5	58.0	14.6
Gdynia	6.0	72.3	5.6
Sopot	3.8	64.5	3.4

Source <https://bdl.stat.gov.pl/BDL>

Based on numbers related to stock in 2019 (shown in Table 1), it is clear that the residential property market in the Tri-City area is quite dynamic compared to the rest of the country. In Gdańsk, there are almost three times more apartments per 1,000 inhabitants compared to all of Poland. On average, flats are smaller in the Tri-City area than in the rest of Poland.

In the primary market, average sale prices of properties larger than 80 square metres (sqm) are above the regional average. The average price of such properties in the primary market is higher than the general average of 7,500 PLN per sqm. In the secondary market, the situation is reversed. Price per sqm is higher for smaller properties than for larger ones. It is more profitable to buy large apartments over 80 sqm from the primary market or flats smaller than 80 sqm from the secondary market. Property prices in Sopot show a trend similar to Gdańsk. There is a significant price difference between the areas preferred by tourists and the non-tourist areas (for example, residential properties closer to the seaside are more expensive). It is challenging to find new investments due to the lack of space to build new residential properties. Due to limited stock and high demand in the primary market, average price per sqm is higher than average price per sqm in the secondary market (Jasińska, 2019).

Gdańsk, Gdynia, and Sopot vary substantially in the number of available residential buildings: in 2019, the number for Gdansk was 30,410, for Gdynia 18,332, and for Sopot 2,779. Sopot is the smallest city and is located between the other two. Gdańsk is the oldest and the largest in Tri-City. Gdynia was recognized as a city in 1926. Thanks to a newly built seaport, Gdynia has grown quite rapidly. The regional railway SKM helps residents of Tri-City travel comfortably through the cities with one pass card or ticket type, and a motorway connects the cities. The residential market continues to grow rapidly in the area: last year 6,221 new apartment buildings were completed in Gdańsk, 1,057 in Gdynia, and 11 in Sopot (Fig. 1).

Most of the transactions in both secondary and primary markets pertained to middle-sized flats in Tri-City. In Gdańsk, both markets are substantially impacted by small flats (around 26% of the market). In contrast, in Gdynia almost half of the primary market consists of middle- and large-sized apartments (29 and 19%, respectively). Interestingly, Sopot's secondary market seems to be the most balanced: the largest apartments represent 19% of all transactions, compared to 7% in Gdańsk and 11% in Gdynia.

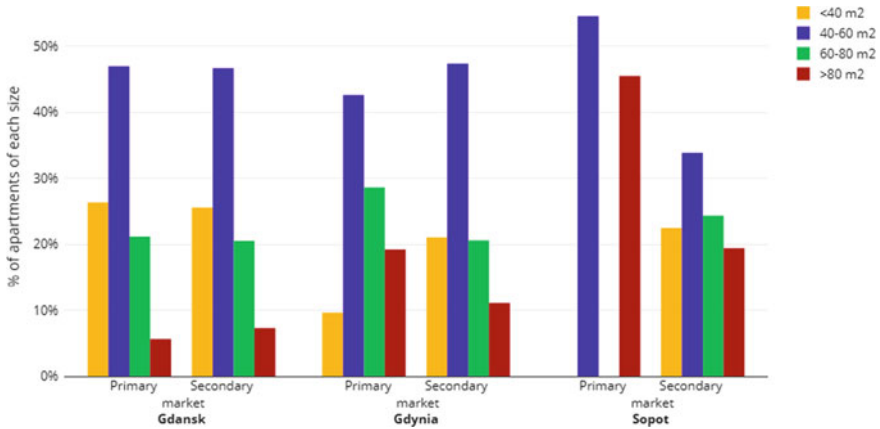


Fig. 1 Distribution of size of apartments in Tri-City Source <https://bdl.stat.gov.pl/BDL>

### 3 Literature Review

Various methods are used to analyse the residential market. Commonly used ones are clustering methods, which help analyse the market based on various attributes. Self-organizing maps (SOM) are not as popular as traditional methods commonly described in the literature, such as k-means. However, there are some studies on the usage of SOM for analysing the real estate market. Eero Carlson applied SOM to data obtained from the National Land Survey of Finland to analyse the survey results (Carlson, 1998). Vilinus Kontrimas and Antanas Verikas compared the performance of SOM with OLS linear regression, support vector machine, and multilayer perceptron for appraisal of properties in Lithuania (Kontrimas & Verikas, 2011). In addition to SOM, k-means combined with geostatistical methods were used to predict prices in each cluster, separately, for Siedlce, Poland (Calka, 2019). A study conducted in 2015 focussed on comparing areas in the centre and suburbs of Shanghai in terms of prices and public infrastructure (Li, 2015). In a study for the residential property market in Turkey, hierarchical clustering algorithms were used to determine homogenous groups of portfolios (Hepsen, 2011).

Poland has a diverse residential property market. In 2014, Łukasz Mach used clustering techniques to create homogenous groups in the real estate markets in voivodeships using both economic and demographic indicators (Mach, 2014). In a study conducted by Beresewicz, the representativeness of Internet data sources for Poland's real estate market were assessed (Beręsewicz, 2015). The study shows that online sources are essential for small- and medium-sized apartments, whereas for large-sized properties, such sources cannot be considered reliable. As primary and secondary residential markets tend to differ significantly, it is beneficial to analyse them separately. One of the main differences found in the literature is that prices in the primary market are more dynamic than those in the secondary market (Leszczyński,

2017). Źróbek et al. studied the influence of environmental factors on people's decisions when buying apartments. According to their results, price, neighbourhood, and security impact decisions to buy an apartment (Źróbek, 2015). Szopińska et al. conducted a study to find whether there is a correlation between price and the noise level in neighbourhoods (Szopińska, 2013). There are also studies that investigated factors affecting real estate prices (Eliasson, 2010; Dittmann, 2013; Grabkowska, 2016; Włodarczyk, 2003).

## 4 Self-Organizing Map

The self-organizing map is based on an algorithm proposed by Teuvo Kohonen in the beginning of the 1980s. The algorithm, an unsupervised neural network, is known for its versatility in data analysis. SOM's straightforward definition helps researchers easily implement the algorithm. Multiple visualization possibilities offered by the method enable researchers to discover multidimensional datasets visually. Mainly, SOM is used for clustering purposes. It provides a two-dimensional map interpretation for the entire dataset. It is expected that similar data points will be grouped closely on the map.

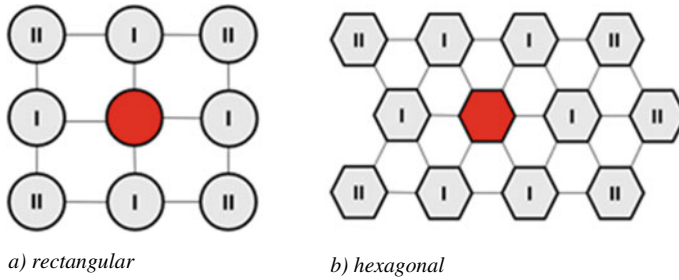
Parameters such as shape and size of the map, topology, neighbourhood radius, functions to measure the distance between observations and neurons, initialization method for the map, learning rate, mode, and number of iterations should be defined by the researchers before moving forward with training the SOM (Kohonen, 1982).

The map's shape and size depend on the space where the neurons are placed. Theoretically, there are no limitations regarding the shape of the map. That being said, unconventional shapes would be challenging to visualize and interpret. The size indicates the distribution of neurons in the space and the number of neurons that will take part in training the SOM. Selecting the wrong size may result in inadequate dataset representation. In case of an extremely small size, complexity will be over-generalized, especially if the number of neurons is less than the number of groups in the dataset. Conversely, making the map's size too large may result in having too many neurons to train, which might lead to noise in the results and longer computation time.

Another parameter impacting SOM structure is topology, which refers to the type of connections within units. Implicitly, topology determines the unit's neighbourhood. The most popular topologies in the literature are rectangular (2–4 connections) and hexagonal (2–6 connections), as shown in Fig. 2.

Units can be linked directly or indirectly by neighbours. The neighbour's rank can be interpreted as the shortest way to get from unit A to unit B using connections with direct neighbours.

Units can be represented in two ways, considering their positions on a map and their weights. The first approach is critical to visualization as it impacts how the results will be understood. Weights are iteratively updated during the training. For example, assume input matrix  $X$  composed of observations  $x_i \in R^k, x_i \in R^k$ . First,



**Fig. 2** Topologies with the level of the neighbourhood of units (a) rectangular, (b) hexagonal

the vector of  $k$ -dimensional weights  $w_m = [w_{m1}, w_{m2}, \dots, w_{mk}]$  representing nodes is initialized and positioned on the dataset’s space. Then, weights are updated to move nodes into the densest parts of the space.

After weights are established, observations are sampled at each iteration, and the algorithm calculates the distance between  $x_i$  and all other weights in  $w_m$ . Any metric can be used for calculating distances. However, Euclidean or Manhattan distance metrics are commonly used by researchers. Best matching unit (BMU, winning neuron) is the neuron with the shortest distance to an observation.

$$d(x_i, w_c) = d(x_i, w_m) \tag{1}$$

After the BMU is obtained, the weights of the neurons are updated. There are two widely used learning types: winner takes all (WTA) and winner takes most (WTM). In WTA, only BMU’s weights are updated. In WTM, weights are adjusted for more than one node. Overall, weights are adapted as follows:

$$w_m(t + 1) = w_m(t) + a(t)G(R, d(c, m))[x(t) - w_m(t)]. \tag{2}$$

where  $t$  is iteration number,  $a(t)$  is learning rate,  $G$  is neighbourhood function, and  $R$  is the neighbourhood’s radius. In WTA training,  $G(R, d(c, m))G(R, d(c, m)) = 1$ . The learning rate impacts the updating of weights. It is suggested to have a higher learning rate at the beginning of training to push neurons faster towards the denser parts of the space.

All neurons in radius  $R$  are updated equally in the rectangular neighbourhood function. In contrast, in the Gaussian neighbourhood function, the closer a neuron is to BMU, the higher its weight. This can be denoted as follows:

$$G(R, d(c, m)) = \exp\left(-\frac{d^2(c, m)}{2R^2}\right) \tag{3}$$

Accordingly, the radius should also be larger at the first iterations. That way, SOM can quickly adapt to the dataset’s overall profile.

The process of adjusting the weights is repeated until the maximum number of iterations is reached or any other stopping mechanism is triggered (Migdał-Najman, 2013). In batch training, a given number of observations is processed in parallel, and weight adjustment is made after all the calculations.

After training, SOM's performance should be measured. In addition to visual assessment, multiple metrics can help empirically evaluate the performance of the SOM. Mean quantification error (MQE) is used to measure the distance between observations and BMUs. Large values may indicate a too small map size or not enough iterations. Topographical error (TE) determines the proper order of the nodes.

$$MQE = \frac{\sum_{i=1}^n d(x_i, w_c)}{n} \quad MQE = \frac{\sum_{i=1}^n d(x_i, w_c)}{n} \quad (4)$$

$$TE = \frac{\sum_{i=1}^n l(x_i)}{n} \quad TE = \frac{\sum_{i=1}^n l(x_i)}{n} \quad (5)$$

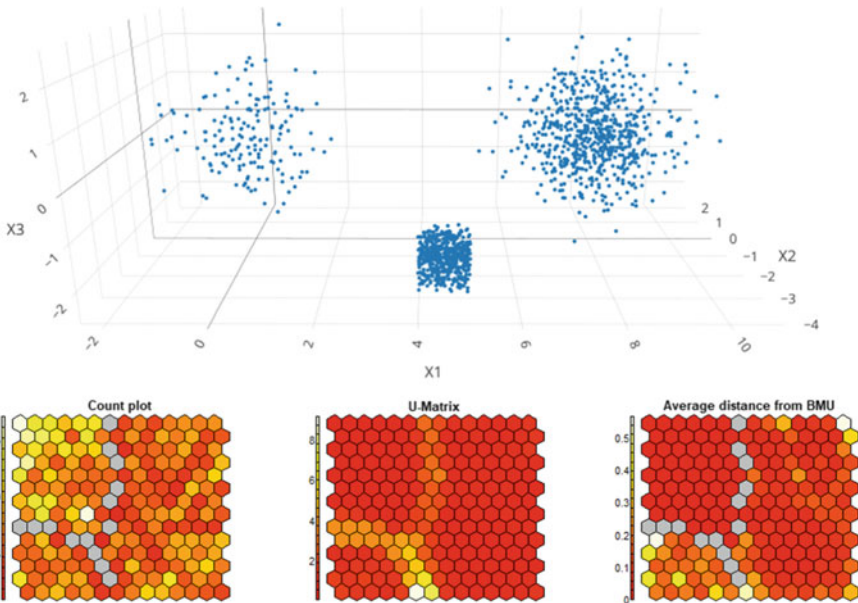
One of the outstanding features of SOM is that it allows researchers to visualize the results. Plots show the mean distance between observations and BMU for each node. In contrast, the frequency map shows how many neurons were selected as winning neurons. U-matrix and frequency maps are analysed together to understand group structures in the dataset (Kohonen, 2012).

As an example, a dataset was created using random numbers to explain SOM practically. For this purpose, a dataset containing three variables X1, X2, and X3 was created. In Fig. 3, the dataset and its related SOM maps are presented. As seen in the plot at the top of Fig. 3, each variable in the dataset has a different distribution.

The U-matrix represented in Fig. 3 shows three different regions full of red hexagons, separated by orange, yellow, and light-red hexagons. The BMU plot shown in Fig. 3 has a similar pattern to the U-Matrix. Grey hexagons split the dataset into three groups on U-Matrix. According to SOM, the dataset can be split into three groups; in other words, three clusters can be created using the dataset.

## 5 Analysing the Residential Market in Tri-City

The R package rvest, which is used for web scraping, was employed to get the most recent listings from the portal <https://ogloszenia.trojmiasto.pl/>, an online portal for real estate listings in Tri-City. The portal provides listings with multiple attributes, such as



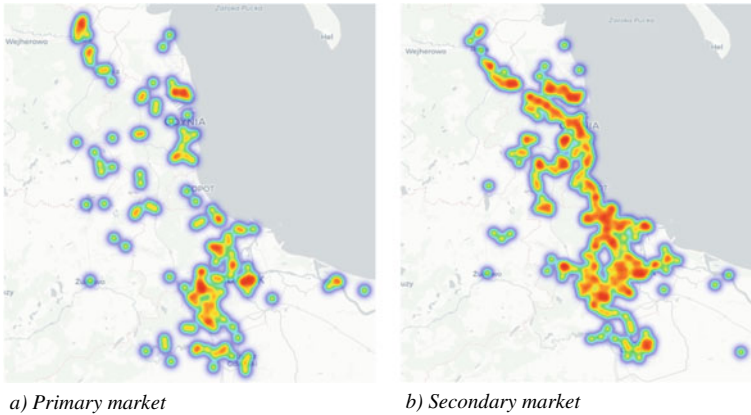
**Fig. 3** Dataset and its SOM representation

- Price of residential property.
- Location of residential property.
- Size of residential property.
- Building age, estimated date of completion (for primary market).
- Property type, apartment or house.
- Total floors in the building.
- Number of rooms in a residential property.
- Creation date, when a listing was created on the portal.

The analysis was performed separately for the primary and secondary markets. Geographical locations of properties obtained from Trojmiasto.pl were converted into geographical coordinates based on latitude and longitude. As it is easier for SOM to handle numeric attributes as opposed to multiple one-hot-encoded ones, coordinates are used instead of city and street names. Each numeric attribute was normalized in  $[0, 1]$  range to smooth out the feature's impact.

For the primary market, 1,231 listings were used in the analysis. As the heat maps presented in Fig. 4 show, the geographical distribution of the apartments differed per market. Newly built properties are located mostly outside of Tri-City, in towns such as Rumia, Reda, and Pruszcz Gdański. There is also a considerable market in the southern part of Gdańsk (Jasień, Kowale, Borkowo) and in the centre of the city (Śródmieście, Zaroślak). For Gdynia, key areas are the northern side of the seaport (Obłuże) and the area close to Sopot (Karwiny, Mały Kack). In Sopot, there are few listings for newly built residential properties.





**Fig. 4** Heatmap of residential property listings in Tri-City, (a) Primary market, (b) Secondary market

A total of 1,317 listings were gathered for the secondary market. It is important to note that there were 354 listings for different property types in the secondary market, such as land, agricultural parcels, warehouses, parking places, and gardens. Because the study’s scope is only apartments, houses, and villas, other listings were removed from the dataset, and they were not included in the heatmap. As shown in Fig. 4b, there are more listings available in the locations closest to the sea.

The R package kohonen (Wehrens, 2007) was used to apply the SOM, which was trained for the primary market with grid size (8,16) and hexagonal topography. The learning rate was updated linearly in the range (0.8, 0.01). After 10,000 iterations, the MQE was 0.008 and TE was 0.2. SOM maps were created to determine correlations between features.

The plots shown in Fig. 5 represent SOM maps. Each plot shows a map for each feature in the dataset. Hexagons on the plots represent neurons, while the colour of each neuron indicates the value of the corresponding feature in the weight vector. Looking at the maps simultaneously helps to understand correlations between features. As shown in Fig. 5, the SOM map ‘Size’ has only one white node, located on the third row from the bottom and eighth column from the left side. According to the colour scale, which is shown on the left side of each map, white colour indicates the highest observations, meaning that the white node represents residential properties with the largest sizes. On the maps ‘Price’ and ‘Number of Rooms’, the white node is located at the same coordinates as the white node on the map ‘Size’. This result shows that the larger the apartments are, the more rooms they have and the more expensive they are. Figure 6 presents geographical maps of the properties grouped inside of each node. The colour of the nodes represents the price range, while the size of the nodes represents the number of observations assigned to the nodes. As shown in Fig. 6a, there are fewer nodes located on the outskirts of Tri-City. However, these nodes have more observations than other nodes located in the centre of Tri-City.



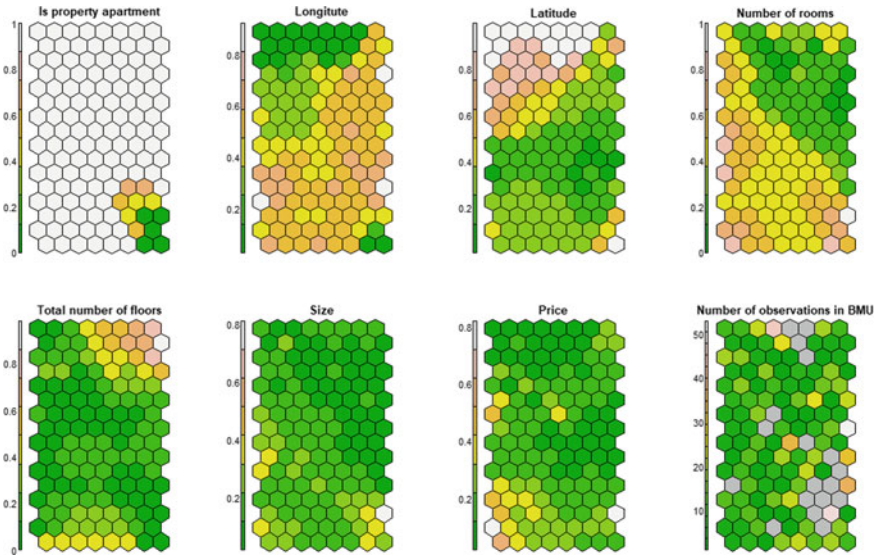


Fig. 5 Visualization of SOM for the primary market

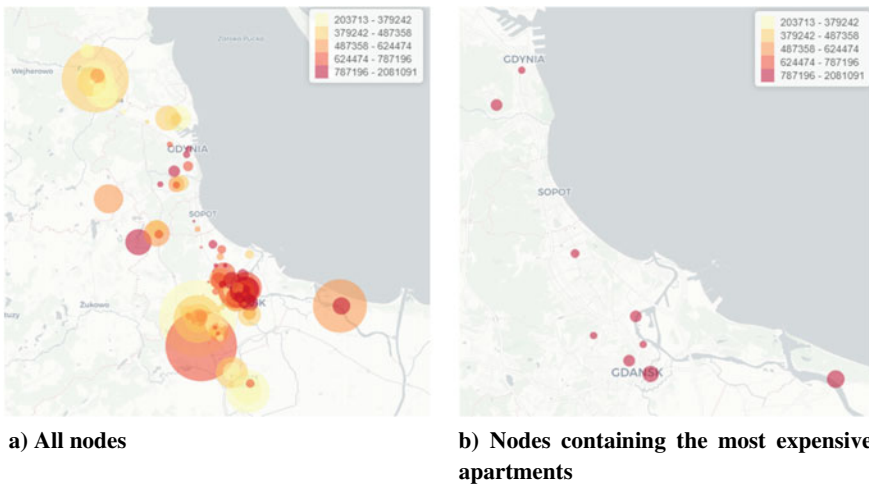


Fig. 6 Geographical map of nodes for the primary market, (a) All nodes, (b) Nodes containing the most expensive apartments

This finding shows that there is more stock in areas such as Reda, Rumia, Jasiień, Kowale, and Pruszcz Gdański. For example, in Reda and Rumia, the price of apartments varies from 203,713 PLN to 624,474 PLN. Although the stock is high in this area, the price range is from 203,713 PLN to 379,242 PLN for most apartments.

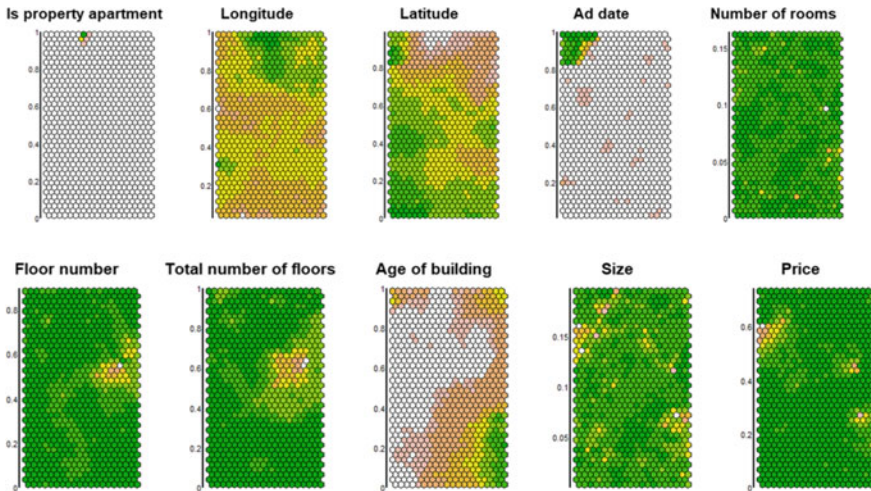


Fig. 7 Visualization of SOM for the secondary market

In Kowale, apartment prices vary from 203,713 PLN to 2,081,091 PLN; however, for most apartments in this area, the price range is from 787,196 PLN to 2,081,091 PLN. Figure 6b shows that the most expensive properties are either near the regional railway lines, SKM and PKM, or in Wyspa Sobieszewska, which is becoming popular as a tourism destination.

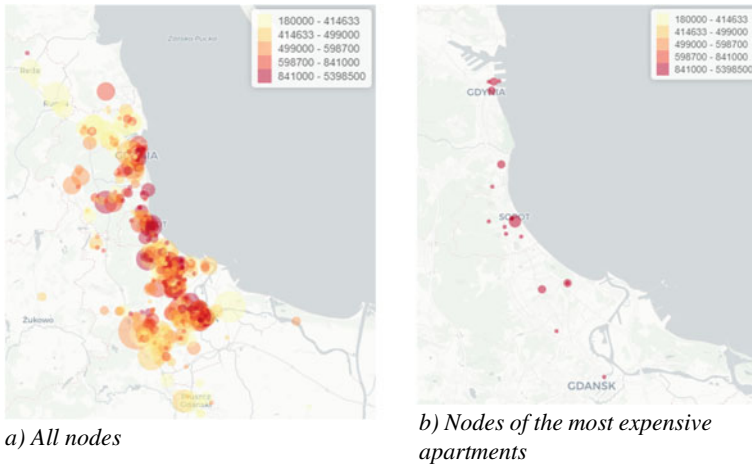
In order to analyse the secondary market, SOM was trained with size (20,40), hexagonal topology, the sum of squares as the distance metric, and the learning rate updated linearly in range (0.8, 0.01). After 10,000 iterations, MQE was equal to 0.002 and TE was 0.25. Additional attributes such as listing date, building age, and floor number were used.

According to the SOM maps shown in Fig. 7, the prices of properties are strongly connected to their size.

The top-left corner of the map ‘Ad Date’ represents listings available on the market for a longer time. More complicated connections between various data are represented on the same parts of the maps.

For example, as depicted in the upper-left clusters of individual maps, older and smaller properties located in taller buildings are relatively cheap, but they have nonetheless been available on the market for a long period. Thus, the conclusion is that these apartments are more difficult to sell.

Figure 8 represents geographical maps for the properties grouped inside nodes. The colours of the nodes refer to the price range, while the size of the nodes refer to the number of observations assigned to the nodes. As shown in Fig. 8a, proximity to the seacoast and the SKM railway has a significant impact on prices. Gdańsk Centre and Przymorze show heterogeneous distribution in terms of prices. Prices range from 499,000 PLN to 596,700 PLN in Kowale, whereas prices in Borkowo do not exceed 499,000 PLN. In Reda, Rumia, and Pruszcz, apartment prices are at the



**Fig. 8** Location of nodes for secondary market data, (a) All nodes, (b) Nodes of the most expensive apartments

lowest level in the market. As presented in Fig. 8b, the most expensive apartments (above 841,000 PLN) are located in Gdynia Centre and Sopot, and there are some exclusive flats in Gdańsk Oliwa and Orunia. The largest node with the highest prices is located in Sopot. Moreover, Fig. 8a shows that there are unlikely to be high-priced properties outside of the city centre.

## 6 Conclusion

In this study, a self-organizing map was used to analyse approximately 2,500 residential property listings in Tri-City's primary and secondary markets. This method proved to be a productive way of analysing and forecasting the real estate market. The algorithm correctly detected correlations between price, size, location, building age, and age of listing. This will allow researchers to draw conclusions and create visuals; it will also help them share results with professionals who do not have data analysis background.

The analysis showed that a property's location significantly impacts its price. In the primary market, location has the most significant impact on price, and in the secondary market, a property's location and size have a notable effect on price. Results revealed that the market in the outskirts of Tri-City is more homogenous in terms of price distribution and attributes of the properties. When it comes to flats in the city centre, it was found that the closer the properties are to the city centre, the more expensive they are. Due to a lack of free space to build new residential buildings in the city centre, fewer properties are listed for sale. The floor number of the properties does not impact the price significantly unless it is relatively high.

Flats on the 13th floor and higher are more expensive than those on lower floors. The algorithm also detected that apartments which are smaller, older, and located in taller buildings are available on the market for a longer time. In other words, it is more difficult to sell them, even though the average price of these apartments is lower than the market average. This correlates with the history of the residential market in Tri-City, which is well known to local people but can be unfamiliar to outside investors. This demonstrates the algorithm's sensitivity as building age itself is not a factor that affects price. Standard and luxury properties in variously aged buildings can be found on the market. Smaller, older apartments in taller buildings are specific to Poland's housing standards of the post-World War II period.

The algorithm also reveals the way a city expands. In this case, results show that Tri-City's primary market expands towards Reda, Pruszcz Gdański, Wyspa Sobieszewska, and Rumia. This determination can provide helpful insight for investors.

## References

- Aydin O, Hayat EA (2018) Estimation of housing demand with Adaptive Neuro-Fuzzy Inference Systems (ANFIS). The impact of globalization on international finance and accounting. Springer, Cham, pp 449–455
- Beręsewicz ME (2015) On representativeness of internet data sources for real estate market in Poland. *AJS* 44(2):45–57. <https://doi.org/10.17713/ajs.v44i2.79>
- Carlson E (1998) Real estate investment appraisal of land properties using SOM. pp 117–127. [https://doi.org/10.1007/978-1-4471-3913-3\\_8](https://doi.org/10.1007/978-1-4471-3913-3_8)
- Calka B (2019) Estimating residential property values on the basis of clustering and geostatistics. *Geosciences* 9(3):143. <https://doi.org/10.3390/geosciences9030143>
- Dittmann I (2013) Primary and secondary residential real estate markets in Poland—analogy in offer and transaction price development. *R Estate Manag Valuat* 21(1):39–48. <https://doi.org/10.2478/remav-2013-0006>
- Eliasson J (2010) The influence of accessibility on residential location. pp 137–164. [https://doi.org/10.1007/978-3-642-12788-5\\_7](https://doi.org/10.1007/978-3-642-12788-5_7)
- Erpolat Tasabat S, Aydin O, Hepsen A (2018) Prediction of residential gross yields by using a deep learning method on large scale data processing framework. *J Bus Econ Financ* 7(1):125–130. <https://doi.org/10.17261/pressacademia.2018.801>
- Grabkowska M, Frankowski J (2016) Close to the city centre, close to the university. Are there symptoms of studentification in Gdańsk, Poland? *32(32):73–83*. <https://doi.org/10.1515/bog-2016-0016>
- Hepsen A, Aydin O, Vatandas O (2015) K - Ortalama Algoritması ile Kümelenmiş Konut Fiyatlarının Fonksiyonel Veri Analizi: İstanbul Örneği. *Finans Politik Ve Ekonomik Yorumlar Dergisi* 52(604):75–85
- Hepsen A, Vatansever M (2011) Using hierarchical clustering algorithms for Turkish residential market. *IJEF* 4(1). <https://doi.org/10.5539/ijef.v4n1p138>
- Jasińska E (2019) Impact of environmental and climate conditions on the investment potential of real estate in the belt of the Gulf of Gdansk Coast. *E3S Web Conferences*, vol 86, pp 00013. <https://doi.org/10.1051/e3sconf/20198600013>
- Karta Trójmiasta podpisana (2007) Retrieved January 20, 2021, from <https://trojmiasto.wyborcza.pl/Tr%C3%B3jmiasto/1,35612,4020355.html?disableRedirects=true>

- Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43(1):59–69. <https://doi.org/10.1007/bf00337288>
- Kohonen T (2012) *Self-organizing maps*. Springer Science & Business Media
- Kontrimas V, Verikas A (2011) The mass appraisal of the real estate by computational intelligence. *Appl Soft Comput* 11(1):443–448. <https://doi.org/10.1016/j.asoc.2009.12.003>
- Li H, Wang Q, Shi W, Deng Z, Wang H (2015) Residential clustering and spatial access to public services in Shanghai. *Habitat Int* 46:119–129. <https://doi.org/10.1016/j.habitatint.2014.11.003>
- Leszczyński R, Olszewski K (2017) An analysis of the primary and secondary housing market in Poland: evidence from the 17 largest cities. *Balt J Econ* 17(2):136–151. <https://doi.org/10.1080/1406099x.2017.1344482>
- Mach Ł (2014) Próba budowy homogenicznych grup województw w obszarze lokalnych rynków nieruchomości mieszkaniowych. *Metody Ilościowe w Badaniach Ekonomicznych* 15(3):219–227
- Migdał-Najman K, Najman K (2013) Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych. *Teoria i zastosowania w ekonomii*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk
- Szopińska K, Krajewska M (2013) Prices of apartments in relation to noise level in Poland. *J Civ Eng Arch* 7(10). <https://doi.org/10.17265/1934-7359/2013.10.001>
- Wehrens R, Buydens LM (2007) Self- and super-organizing maps in R: the Kohonen package. *J Stat Softw* 21(5):1–19
- Włodarczyk D, Dymnicka M (2003) 1. How do people want to live? : Residential preferences and values in Gdynia, Poland. In *Building and Re-building sustainable communities : Reports from the Superbs project*, 1st edn, pp 4–12
- Żróbek S, Trojanek M, Żróbek-Sokolnik A, Trojanek R (2015) The influence of environmental factors on property buyers' choice of residential location in Poland. *J Int Stud* 7(3):163–173

# Advanced Car Price Modelling and Prediction



Michail Tsagris and Stefanos Fafalios

**Abstract** The scope of the paper is modelling and prediction of brand new car prices in the Greek market. At first the most important car characteristics are detected via a state-of-the-art machine learning variable selection algorithm. Statistical (log-normal regression) and machine learning algorithms (random forest and support vector regression) operating on the selected characteristics evaluate the predictive performance in multiple predictive aspects. The overall analysis is mainly beneficiary for consumers as it reveals the important car characteristics associated with car prices. Further, the optimal predictive model achieves high predictability levels and provides evidence for a car being over or under-priced.

**Keywords** Car market · Price prediction · Variable selection · Nonlinear models

## 1 Introduction

Car price modelling and prediction has not attracted significant research interest, especially in the field of economics, the most suitable environment for research in this area. Some examples include Eckard (1985) who showed, empirically, that U.S. state regulation of new car dealer entry produces higher new car prices, as predicted by economic theory. Verboven (1996) explained the presence of price discrimination across European countries and attributed this phenomenon to cross-country differences in price elasticities, differences in quota regimes and differences in the degree of collusive behaviour. Matas (2009) constructed a quality-adjusted price index for the Spanish car market over the period 1981–2005. However, those papers cannot be exploited for the present analysis, due to the rapid technological advances and the adoption of sophisticated functionalities such as ESP, parking and weather sensors not available at that time.

---

M. Tsagris (✉)

Department of Economics, University of Crete, Gallos Campus, Rethymnon, Crete, Greece  
e-mail: [mtsagris@uoc.gr](mailto:mtsagris@uoc.gr)

S. Fafalios

Gnosis Data Analysis, Herakleion, Crete, Greece

More recently, Busse et al. (2013) compared the relationship between prices of gasoline to prices of used and new cars, and Gegic et al. (2019) predicted the discretized, into mutually exclusive classes, car prices. Xia et al. (2020) predicted the car sales using a highly versatile machine learning algorithm and Alberini et al. (2016) conducted, in the Swiss market, an analysis that resembles to some extent the current analysis. Aiming at examining whether fuel economy is capitalized in the car price, they linked price to car characteristics, collecting panel data but without considering key features, such as brands.

Wu et al. (2009) proposed an expert system to forecast the price of used cars using an adaptive neuro-fuzzy inference system. Lessmann and Voss (2017) empirically investigated numerous linear and nonlinear statistical models for forecasting the resale prices of used cars. Andrews and Benzing (2007) analysed the influence of auction, seller and product on the price premium in an eBay used car auction market. On a different route, Raviv (2006) examined the sequence of winning bids in the public auction of used cars in New Jersey providing evidence of an order-dependent increase in the price.

The current paper combines the concepts of the more recent work and attempts to extend it by considering the Greek car retail market in December 2020. Its scope is oriented towards the prediction of the prices of brand new cars. Specifically, using information from the characteristics of new cars the goal of the paper is to accurately model and predict the car prices. This information is mainly of importance for consumers who want to know the impact of each car characteristic on its price and whether additional characteristics and associated costs translate to true consumer value. To this end, a selection of the important car characteristics affecting its pricing is initially performed. Using the selected characteristics, statistical and machine learning algorithms yield interesting conclusions regarding the effect of those characteristics on the prices. Machine learning algorithms proved useful in terms of predictive performance while further examination of the final model's predictability pointed out the consumer benefits by providing strong evidence as to which cars are estimated to be over or under-priced.

The rest of the paper is organised as follows. The data analysed are described followed by a delineation of the models and algorithms whose predictive capability is assessed in multiple directions. Finally, conclusions close the paper.

## 2 Data Description

Cross-sectional data on brand new cars were accessed from the popular Greek car site [autotrivi](http://autotrivi.gr) in December 2020 covering a total of 1,600 brand new cars on 39 characteristics (price, horsepower, engine displacement, fuel type, time to 100km/h, fuel consumption, etc.). Missing data information mandated a pre-processing prior to the analysis.



## 2.1 Data Cleaning Process

The car prices ranged from €9,100 up to €283,760 with mean and median values equal to €36,796 and €27,680 respectively. To safeguard against the high variability and in order to make safer predictions, only cars priced under €50,000 were selected. According to the [World Bank](#) the estimated Greek GDP per capita for 2019 was \$19,582 justifying the choice of our selection.

A further investigation emerged the necessity to remove more cars. Three additional cars in the current dataset operating with LPG were excluded. This initial “cleaning” process divulged that Alfa Romeo should participate with only 9 car models, Mitsubishi with 9 models, and Land Rover and Lexus with 1 model each. Since these brands had less than 10 models and hence carry little information they were removed from further analysis. The reason being is the tenfold cross-validation protocol, where the split of the data took place in a stratified manner ensuring that each fold contained all brands. These actions along with the removal of cars with missing information on their characteristics ensured that 909 cars (models) would participate in the analysis, a number high enough to perform valid statistical inference and draw reliable conclusions. The 39 car characteristics of these 909 cars appear in [Table 1](#), while the type of fuel and category appear on the contingency [Table 2](#).

## 2.2 Brands and Car Prices

The ranges of the car prices grouped by the brand are visualised in [Fig. 1](#). BMW, Subaru, Mercedes and Volvo sell the most expensive cars, all above €20,000, whereas Dacia, Fiat and Suzuki are the lowest priced cars, with no car over €30,000. The two most expensive cars belong to BMW, both valued more than €49,000, whereas the lowest-priced car is manufactured by Seat, valued €9,100. It is worth highlighting that Dacia and Fiat trade cars are also priced below €10,000. Nissan, Mazda, Audi and Honda produce cars at a wide range of prices, whereas the range of prices of Dacia, Fiat and Citroen is relatively small, compared to the other brands. Evidently, the car prices differ significantly across the brands and no statistics are necessary to validate this.

According to the overall number of sales in the Greek market,<sup>1</sup> Toyota is the most popular brand, whereas Subaru is the least popular brand. [Figure 1](#) manifested that Toyota and Peugeot sell medium-priced cars, yet these two brands hold the highest number of sales. Fiat and Dacia on the other hand sell the most economic cars, yet they acquired the 15th and 16th position in the number of sales, respectively. The Spearman correlation between the brands ranked according to their median car prices and ranked according to their number of sales is equal to  $-0.291$ . This manifests a

---

<sup>1</sup> Information accesses through [autotriti](#).

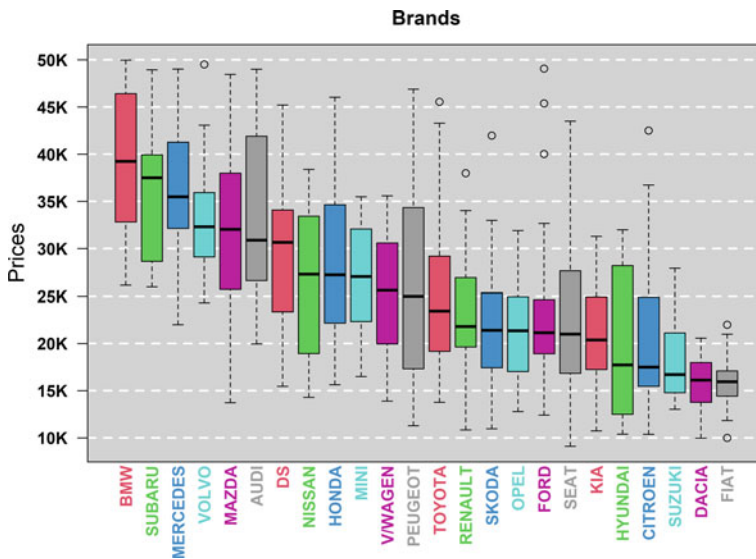


**Table 1** Car characteristics

Characteristic	Mean	Median	Minimum	Maximum	Characteristic	Yes	No
Engine (cc)	1,454	1,496	898.00	2,494	Air-condition	293	616
Time to 100km/h (s)	10.41	10.30	5.30	17.10	Clima	649	260
Consumption (L/100 km)	5.00	5.00	3.00	8.10	Rear electric windows	784	125
CO <sub>2</sub> emissions (g/km)	118.20	116.00	76.00	189.00	Fog lights	795	114
Reservoir autonomy (km)	1,020	980.00	593.00	1,842	Parking sensors	645	264
Taxation (€)	138.40	114.00	0.00	525.00	Rain sensors	572	337
Length (mm)	4,311	4,363	3,466	4,871	Xenon lights	225	684
Width (mm)	1,798	1,800	1,595	1,969	Cruise control	757	152
Height (mm)	1,529	1,495	1,353	1,801	Leather lining	207	702
Distance between wheels (mm)	2,634	2,649	2,300	2,920	Light alloy wheels	768	141
Port baggage size (L)	424.60	400.00	170.00	780.00	Sunroof	217	692
Fuel tank size (L)	49.71	50.00	32.00	70.00	Navigator	383	526
Weight (kg)	1,314	1,325	835.00	1,836	Manual gear box	684	225
Horsepower	132.70	122.00	60.00	1,603			
Rounds/m at maximum hp	4,872	5,000	400.00	6,600			
Torque (Nm)	230.50	240.00	91.00	445.00			
Rounds/m at max (Nm)	2,099	1,750	0.00	5,000			
Cylinders	3.67	4.00	3.00	4.00			
Maximum speed	194.10	193.00	150.00	250.00			
Guarantee in mechanics (years)	4.34	5.00	2.00	8.00			
Guarantee in rust (years)	12.18	12.00	6.00	30.00			
Guarantee in colour (years)	2.97	3.00	2.00	5.00			

**Table 2** Car category and type of fuel

Category	Fuel			Row totals
	Diesel	Gasoline	Hybrid	
Big	14	16	1	31
Small-medium	63	96	17	176
Medium	31	37	0	68
Mini	2	44	8	54
Off-road	136	231	36	403
Polymorph	16	17	0	33
Small	24	114	6	144
Column totals	286	555	68	909



**Fig. 1** Box plot of the car prices across brands ordered according to their median values

negative correlation, which is however non-statistically significant at the 5% level ( $p\text{-value} = 0.260$ ) and implies that brand car prices and the number of sales seem not to be statistically significantly associated.

### 3 Data Analysis

The available 39 car characteristics serve as candidate predictor variables for the logarithm of car prices ( $y$ ). The logarithmic transformation was chosen due to the right skewness of the price distribution.

### 3.1 Selection of the Important Car Characteristics

The task of selecting the important predictor variables (car characteristics) operated under the Forward Backward with Early Dropping (FBED) algorithm<sup>2</sup> (Borboudakis and Tsamardinos 2019). In brief, the algorithm proceeds in a forward manner, while dropping the non-significant predictor variables at each step, attempting to identify the predictor variables that are statistically significantly associated (at the 5% significance level) with the response variable. Upon completion of the forward search, a backward search is applied to remove any falsely selected variables. The FBED algorithm detected 18 car characteristics that are statistically significantly associated with the logarithm of the price.

### 3.2 The Log-Normal Regression Model

The logarithm of the car price implies that a log-normal regression was fitted with the relevant density of the log-normal distribution given by

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right), \quad (1)$$

where  $\mu$  and  $\sigma^2$  refer to the mean and variance parameters of the underlying normal distribution, respectively. Hence, the regression model is of the form  $E(\ln(y)|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ , where  $\mathbf{X}$  and  $\boldsymbol{\beta}$  define the design matrix of the predictor variables and the vector of coefficients, respectively. This model signifies that fitted and predicted car prices based upon the log-normal regression model, when back-transformed to Euros, are equal to  $\exp\left(\hat{y} + 0.5\hat{\sigma}_{y|\mathbf{x}}^2\right)$ , where  $\hat{\sigma}_{y|\mathbf{x}}^2$  is the estimated regression variance, since the mean of the log-normal distribution is  $E(y) = \exp(\mu + 0.5\sigma^2)$ . The logarithmic transformation was not applied to the continuous car characteristics despite their units of measurement being positive so as to keep their effects more interpretable. No interactions among the predictor variables were added either, as this would surge the number of estimated parameters.

According to the coefficients of the log-normal regression (not shown here), there is a mixture of car characteristics affecting its pricing, both mechanical and image related. The interpretation of those coefficients is straightforward; each of them refers to the expected percentage-wise price change for a given unit change in the values of each car characteristic, *ceteris paribus*. Overall, the coefficients possess the correct sign and their magnitude was also justified by univariate analyses. In terms of model fit, the log-normal model explains the 94.00%<sup>3</sup> of the logarithm of the price variability, providing good evidence of a highly acceptable model fit.

<sup>2</sup> FBED is publicly available in the *R* package *MXM* (Tsagris and Tsamardinos 2019).

<sup>3</sup> The adjusted coefficient of determination is equal to 93.68%.

However, there are three issues that should be considered. The reported (adjusted) coefficient of determination is not a valid prediction evaluation criterion. Secondly, the influence of the characteristics on the car's price might be far from linear and subsequently the prediction error of the log-normal regression model is not the lowest that can be achieved.

### 3.3 Car Price Prediction

The model's coefficient of determination is substantially high and despite revealing a very satisfactory fit, it cannot provide information on the model's predictability as it was computed on the same data the model was fitted, and hence it overestimates the model's true predictive performance.

A better strategy is to apply the tenfold cross-validation (CV) pipeline. This commences with splitting the dataset into ten mutually exclusive folds or sets in a stratified manner. The cars are randomly assigned to each fold in a stratified manner so that the distribution of the brands is nearly the same in all folds and hence each brand will be represented in each fold. One fold is left aside playing the role of the test set, while the other nine folds are collected in what is termed the training set. In the training set, the FBED algorithm selects the most important variables utilising the log-normal regression model. A predictive model is subsequently built and validated on the test set, i.e. using the values of the selected car characteristics in the test set, the car prices are predicted. Three metrics evaluating the predictive performance were estimated during the tenfold CV procedure. The percentage of variance explained (PVE), the mean absolute error<sup>4</sup> (MAE) and the mean error (ME) per brand are defined as

$$\text{PVE} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2a)$$

$$\text{MAE} = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N} \quad (2b)$$

$$\text{ME} = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)}{N} \quad (2c)$$

respectively, where  $\hat{y}_i$  refers to the predicted values of the test set whose sample size is equal to  $N$ . The PVE can be interpreted as the out of sample coefficient of determination. MAE, on the other hand, is easier to interpret and states that, on average, the predicted car price may deviate by plus or minus a value. Lastly, the ME shows the average direction (or sign) of the errors.

This process is repeated ten times so that all folds have played the role of the test set. Due to inherent variability in the results, the pre-described tenfold CV pipeline

---

<sup>4</sup> MAE serves the purpose of practical interpretation.

is repeated ten times with different generated folds each time<sup>5</sup> and the capability of the predictive model stems from the aggregation of all folds across all repetitions.

### 3.3.1 Machine Learning Predictive Algorithms

Highly versatile machine learning algorithms, such as random forest (RF) and support vector regression (SVR) were employed to assist in obtaining more accurate predictions. The algorithms’ perk is the exploitation of the nonlinear functional relationship between the car characteristics and its price which can result in more accurate predictions. The RF algorithm is built upon creating numerous regression trees, justifying its name. RF relies on “bagging” (bootstrap aggregation) (Breiman 2001) and random selection of features (Ho 1995). The algorithm randomly draws a subset of variables with a bootstrap sample,<sup>6</sup> termed  $\mathbf{X}_b$  and  $Y_b$ , and builds a tree using this subset. The tree discretizes the continuous variables into classes seeking for the optimal split, with the number of splits being a hyper-parameter that requires tuning. The process of randomly selecting variables and bootstrap samples is repeated  $B$  times with the predictions being computed over aggregation of all tree-based predictions  $\hat{y} = \frac{\sum_{b=1}^B f(\mathbf{x}_b)}{B}$ .

SVR is more complex and relies upon the following constrained minimization as described in Meyer et al. (2020)

$$\begin{aligned} \min_{\mathbf{a}, \mathbf{a}^*} \quad & \frac{1}{2} (\mathbf{a} - \mathbf{a}^*)^T \mathbf{Q} (\mathbf{a} - \mathbf{a}^*) + \epsilon \sum_{i=1}^n (a_i + a_i^*) + \sum_{i=1}^n y_i (a_i + a_i^*) \\ \text{s.t.} \quad & 0 \leq a_i, a_i^* \leq C, \quad i = 1, \dots, n \\ & \sum_{i=1}^n (a_i - a_i^*) = 0, \end{aligned}$$

where  $\mathbf{a}, \mathbf{a}^*$  are the vector of parameters to be estimated,  $\epsilon$  is a very small quantity,  $C$  is the cost, a tunable hyper-parameter and  $\mathbf{Q}$  is an  $n \times n$  positive semidefinite matrix with elements  $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ , where  $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$  denotes the kernel matrix. The radial basis function  $\exp(-\gamma|\mathbf{u} - \mathbf{v}|^2)$  was selected as the kernel with  $\gamma$  being a tunable hyper-parameter.

The number of splits of the variables examined in the RF was (1, 3, 5, 10). The cost hyper-parameter in SVR laid hold of ten equidistant values spanning from 0.2 to 2, while the  $\gamma$  parameter also took ten values, equally spread between  $1/d^2$  and  $1/d^{0.5}$ , where  $d$  denotes the number of variables. Within the CV protocol, the predicted car prices were now based upon the RF with 4 hyper-parameter values (four splits), and with SVR using all combinations of the cost and  $\gamma$ , in the car characteristics that were selected by FBED. This results in four sets of predicted car prices for the RF and 100 sets for the SVR.

<sup>5</sup> This avoids “lucky” splits that could yield a high predictive performance.

<sup>6</sup> Sample with replacement, of the same size.

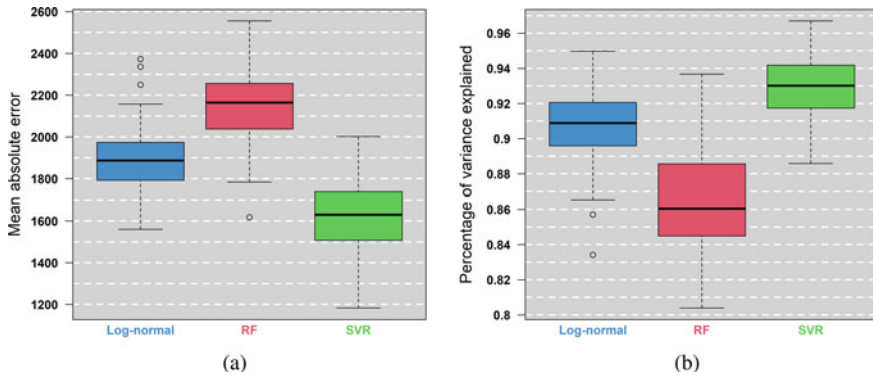


Fig. 2 Performance metrics of the CV protocol: **a** MAE expressed in € and **b** PVE

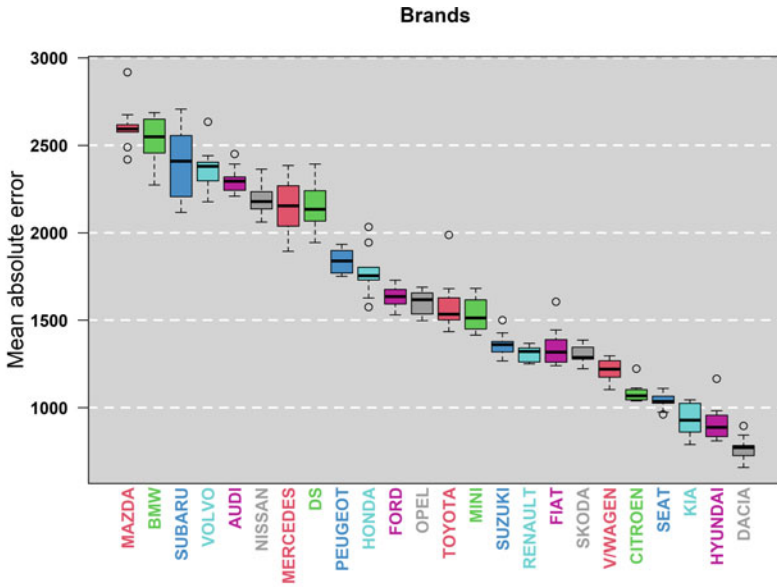
### 3.3.2 Predictive Performance Evaluation Results

The results of the ten times repeated tenfold CV appear in Fig. 2. The RF algorithm yielded the least accurate results, followed by the log-normal regression, while SVR produced the optimal predictions. The estimated PVE (2a) and MAE (2b) values of SVR equal 92.86% and €1,622.19 respectively. MAE, on the other hand, is easier to interpret and states that, on average, the model’s predictions of a car price may deviate by plus or minus €1,622.19. This error is satisfactorily small relative to the range of car prices as it corresponds to 3.97% of the observed range of prices (€9,100–€49,983).

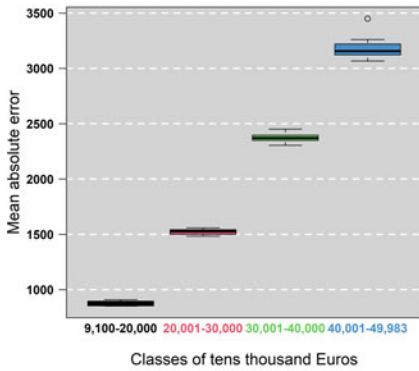
Figure 3 visualises the predictive performance of the SVR. In Fig. 3a MAE is classified according to the brand. MAE differs significantly across the brands and surprisingly enough, the error is high in the cheaper cars, Suzuki and Fiat. However, the highest error was observed for Mazda cars which does come by surprise as this brand produces the fifth most expensive cars (see Fig. 1). What is not surprising though is that the eight most expensive cars are the ones with an MAE more than €2000.

The error increases as the prices increase,<sup>7</sup> as observed in Fig. 3b, where MAE is classified according to intervals of €10,000. This is also evident in the scatter plot of Fig. 3c that visually contrasts the predicted prices against the observed prices, with the blue line corresponding to the 45° line, or perfect agreement. The higher the prices, the higher the spread of their predictions around the blue line, yet the correlation between these two is really high and equal to 0.968.

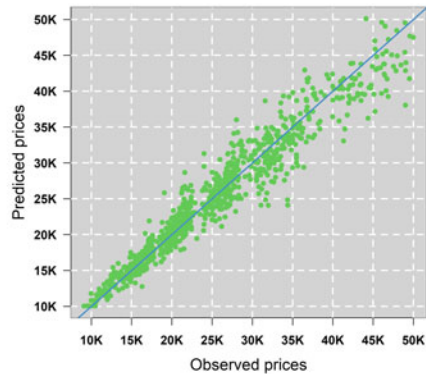
<sup>7</sup> The majority of the cars (71%) are priced less than €30,000 explaining the overall MAE of €1,622.19.



(a)



(b)



(c)

**Fig. 3** Performance metrics of the CV protocol. **a** MAE according to brand, **b** MAE in classes of ten thousand € and **c** observed versus predicted values

### 3.3.3 Estimated over and Under-Priced Cars

From the consumer’s point of view it is also worthy to characterize a car as being over or under-priced, based on the predictions of the cross-validation.<sup>8</sup> Figure 4 displays

<sup>8</sup> To be fair when assessing the over/under-pricing of a car, we had to use the predicted prices and not the fitted prices.

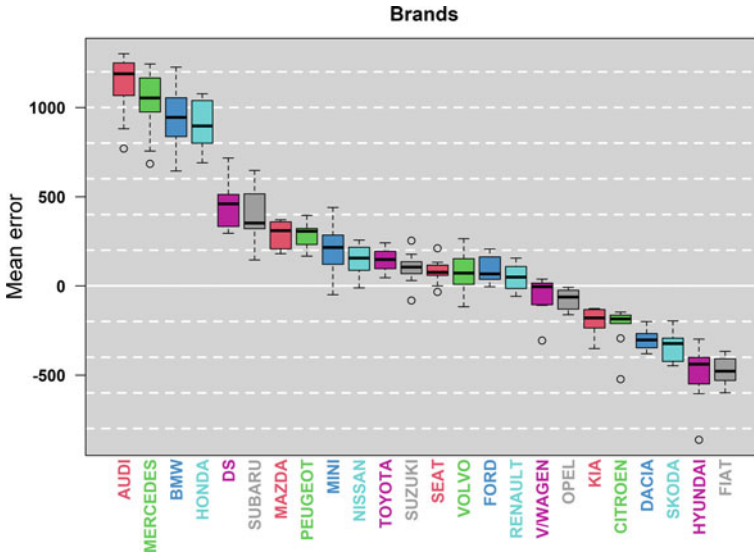


Fig. 4 Performance metrics of the CV protocol II. ME per brand

ME, an informative measure that sheds light into the problem of detecting over and under-priced cars. The boxes of brands located above the zero vertical line indicate brands estimated to be over-priced, whereas boxes of brands located below indicate brands classified as under-priced. Seat is estimated to produce the most over-priced cars, whereas Fiat is estimated to produce the most under-priced cars. Reputation, that is linked to reliability, could be the causal factor attributing to this phenomenon. A suitable proxy for this variable could be the reliability rating index. Even though this index exists, it ought to be country-specific and unfortunately, no values exist for the Greek market.

The largest negative difference between the actual and estimated prices was €11,342.89 observed at an Audi brand. and the largest positive difference was equal to €7,499.14, observed at a BMW branded car. The predicted price of the Audi brand car priced at €49,010 was €37,667.11. By examining the characteristics of cars whose true prices range within a €500 range it is obvious that this is an over priced car. There are 8 cars at a similar range of prices offering the same or better characteristics with 5 of them being produced by the most expensive brands (Volvo, Mazda, Subaru). This implies that a consumer who desires to purchase an Audi car has more affordable options from similar level brands, with similar characteristics. On the other hand, a BMW brand car priced at €31,594 was predicted to be worth €39,093.14. When considering cars priced within a window of €500 far from the predicted value, it is apparent that the characteristics of three cars (Subaru, Toyota and BMW) are similar to that of this BMW car, but at different brands. The referred car, given its high-class brand, can be seen as a value-for-money car. Thus, in this



instance, comparing prestigious cars alone, a BMW and a Subaru cars are more expensive than this BMW car by more than €7,000, despite all three cars having similar characteristics.

### 3.4 *Estimated Individual Effect of the Car Characteristics*

Since SVR does not return coefficients demonstrating the effect of each car characteristic on the price, the individual conditional expectation (ICE) plots (Goldstein et al. 2015) will portray these effects visually. The advantage of these plots is the visualisation of the nonlinear effect of the independent variables on the response variable. In this case study though a bootstrap variant of ICE plots has been implemented, within this bootstrap variant frame a car characteristic is chosen and its values are sampled with replacement. The optimal SVR, corresponding to the hyperparameters that yielded the optimal predictive performance, is fitted. This process is repeated 100 times and the average estimated prices are computed. For the continuous car characteristics the ICE is estimated using a locally-weighted polynomial regression.<sup>9</sup>

Figure 5 shows the effect of each characteristic on the estimated car price. Specifically Fig. 5a–c demonstrate the car brand, fuel and category effect on the estimated prices. The effect of the brand is sorted in descending order. This order is in an almost perfect agreement with the true order of the brands sorted according to their average prices. The only discrepancy is that SVR estimates Ford to be more expensive than Seat, whereas the opposite is true in this sample. As for the fuel the order is as expected, with diesel cars being the most expensive followed by hybrid and gasoline cars. Finally, for the car category the estimated order is distorted only between off-road and polymorphic cars. These plots evidently signify that SVR has managed to detect the correct effect of these categorical valued car characteristics. It must be highlighted that the estimated effects based on the log-normal regression coefficients were not as accurate as the SVR estimated effects.

Figure 5d, e visualises the effect of the continuous car characteristics. Since these characteristics are measured in different units, they were first normalised to be mapped on the same scale using  $\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$ , where  $i = 1, \dots, n$ , with  $n = 909$  cars. The car weight (in kg), acceleration (number of seconds the car requires to reach a speed of 100 km/h), engine displacement (in cc<sup>3</sup>) and width (in mm) are plotted in Fig. 5d. As expected, the acceleration has a negative impact on the car price as the longer the time required for a car to reach the 100 km/h speed, the less expensive it is. The heavier and the bigger (in terms of engine) the car is, the more expensive it is. Car width is positively associated with its price up to a certain point, above which the car width influences price in a negative manner. The car torque (in Nm) and its maximum speed are two characteristics positively associated with the price as depicted in Fig. 5e. On the same graph, it is observed that the size of the port baggage

<sup>9</sup> For the categorical car characteristics, this kernel regression step is omitted.

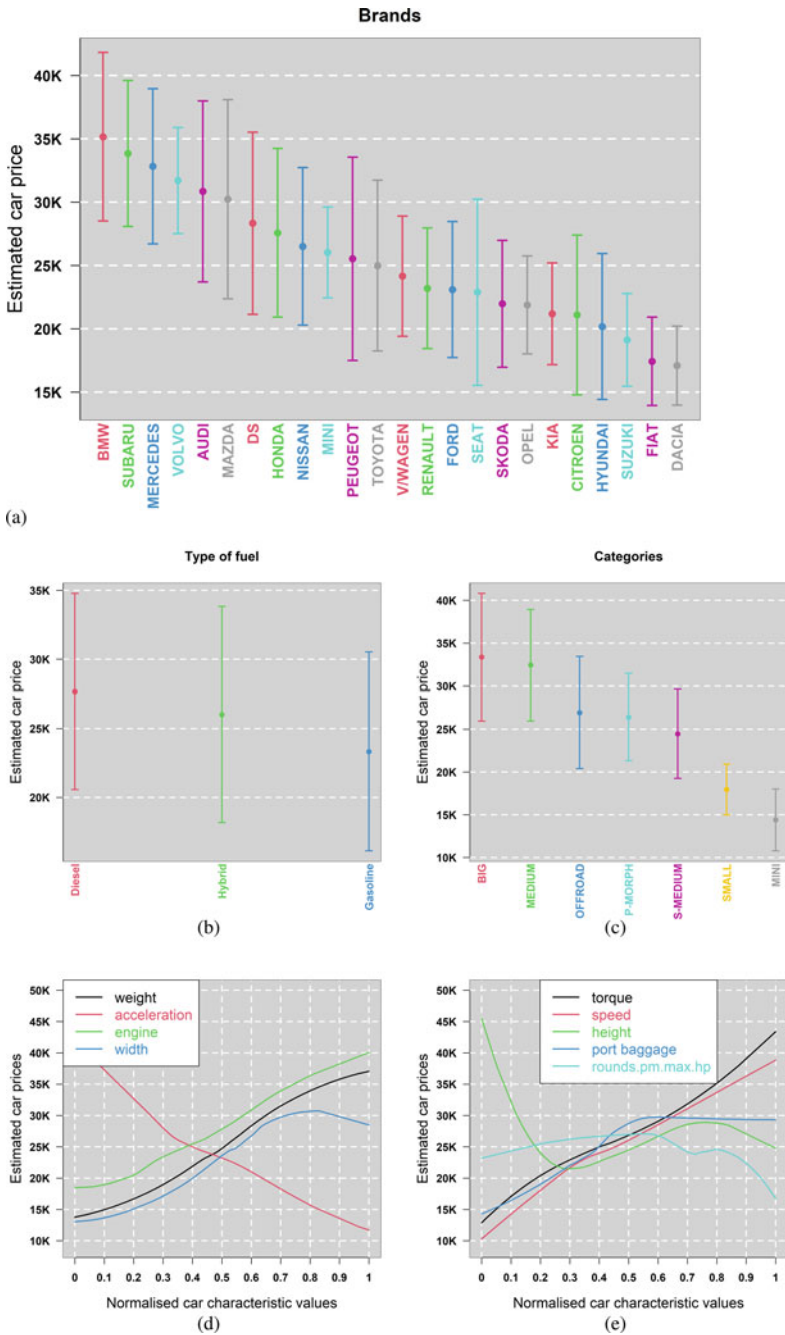


Fig. 5 Effect of each car characteristic on the estimated car price

(in litres) has a partial positive effect up to a certain threshold, after which it becomes flat and hence does not affect the price. The effect of the height is highly nonlinear and a possible explanation could be that shorter cars are perhaps convertibles or sportive. As the height increases the car becomes sedan, hatchback or coupe style, so usually less expensive. A higher car indicates an SUV type or Jeep type which are more expensive. Rounds per minute at the maximum horsepower does not seem to be associated with the price. Note that these are individual effects, so plotting these in higher dimensions could reveal more interesting patterns with regards to their combined effects.

## 4 Conclusions

A variable selection algorithm identified significant associations between car characteristics and pricing in the Greek market. Among the 18 identified characteristics, the most important were the car's weight, brand, time to reach 100 km/h, category, torque, type of fuel and maximum speed. It is natural to assert that the effect of the identified car characteristics differs across brands. Allowing for interactions between car brands and the rest of the variables would relax the rather restrictive assumption imposed on the slopes of the continuous variables. However, the addition of such interaction terms would increase dramatically the number of estimated parameters and negatively affect the validity of the estimates. Such an approach would require either larger car samples or the missing information for all cars to be available on the [autotriti](#)'s website.

Although the model fit with the use of a log-normal distribution was excellent, the predictive ability of the model was sub-optimal. This discrepancy could be due to nonlinear underlining associations between car characteristics and price, an association better captured by using machine learning techniques. Such methods produced models with greater predictive ability than the regression models at the expense of interpretability. The SVR predicted the car prices with a deviation of nearly €1,622 which can serve as a guideline for consumers.

Price prediction that is higher than the actual price indicates evidence of a value-for-money car or evidence of an interesting purchase. On the other hand, a lower than the actual predicted price signals a rather over-priced car whose purchase might not be for the consumer's best of interest. Two extremely over and under-priced cars exhibited this phenomenon. Note, however, that the characterization as over or under-priced relies upon the available data and it could also be attributed to chance as these are estimates. Moreover, there are unobserved factors contributing to the car price, such as research and development costs and marketing/advertisement expenses that are not publicly available. Brand reputation and reliability were not measured either and the records of each brand regarding mechanical faults observed after the purchase cannot be undisclosed to the public.

Collectively, all the aforementioned results are beneficial for consumers and can act as a guide for choosing a value-for-money car. The results further signified that not all machine learning algorithms do necessarily outperform statistical models, yet, some of them can.

If documentation of the production cost of each car was available, application of a stochastic frontier model (Battese and Coelli 1995) would further evaluate the profit efficiency of the car companies. This would enable companies to better orient their strategies and make the whole vehicle market more efficient.

**Acknowledgements** The authors are grateful to Nikolaos Pandis, Eleftherios Pavlos and George Filis who read earlier versions of this manuscript and provided constructive feedback.

## References

- Alberini A, Bareit M, Filippini M (2016) What is the effect of fuel efficiency information on car prices? Evidence from Switzerland. *Energy J* 37
- Andrews T, Benzing C (2007) The determinants of price in internet auctions of used cars. *Atl Econ J* 35:43–57
- Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empir Econ* 20(2):325–332
- Borboudakis G, Tsamardinos I (2019) Forward-backward selection with early dropping. *J Mach Learn Res* 20:1–39
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Busse MR, Knittel CR, Zettelmeyer F (2013) Are consumers myopic? Evidence from new and used car purchases. *Am Econ Rev* 103:220–56
- Eckard EW Jr (1985) The effects of state automobile dealer entry regulation on new car prices. *Econ Inq* 23:223–242
- Gegic E, Isakovic B, Keco D, Masetic Z, Kevric J (2019) Car price prediction using machine learning techniques. *TEM J* 8:113–118
- Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 24:44–65
- Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1, pp 278–282
- Lessmann S, Voss S (2017) Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy. *Int J Forecast* 33:864–877
- Matas A, Raymond JL (2009) Hedonic prices for cars: an application to the Spanish car market, 1981–2005. *Appl Econ* 41:2887–2904
- Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2020) e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4
- Raviv Y (2006) New evidence on price anomalies in sequential auctions: used cars in New Jersey. *J Bus Econ Stat* 24:301–312
- Tsagris M, Tsamardinos I (2019) Feature selection with the R package MXM. *F1000Research* 7:1505

- Verboven F (1996) International price discrimination in the European car market. *RAND J Econ* 27:240–268
- Wu JD, Hsu CC, Chen HC (2009) An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Syst Appl* 36:7809–7817
- Xia Z, Xue S, Wu L, Sun J, Chen Y, Zhang R (2020) ForeXGBoost: passenger car sales prediction based on XGBoost. *Distrib Parallel Databases* 38:713–738

# Impact of Outlier-Adjusted Lee–Carter Model on the Valuation of Life Annuities



Cem Yavrum and A. Sevtap Selcuk-Kestel

**Abstract** Annuity pricing is critical to the insurance companies for their financial liabilities. Companies aim to adjust the prices using a forecasting model that fits best to their historical data, which may have outliers influencing the model. Environmental conditions and extraordinary events such as a weak health system, an outbreak of war, and occurrence of pandemics like Spanish flu or Covid-19 may cause outliers resulting in misvaluation of mortality rates. These outliers should be taken into account to preserve the financial strength and liability of the life insurance industry. In this study, we aim to determine if there is an impact of mortality jumps in annuity pricing. We question the annuity price fluctuations among different countries and two models on country characteristics. Moreover, we show the annuity pricing on a portfolio for a more comprehensive assessment. To achieve this, a simulated diverse portfolio is created for the prices of four types of life annuities. Canada, Japan, and the United Kingdom as developed countries with high longevity risk, Russia and Bulgaria as emerging countries are considered. The results of this study prove the use of outlier-adjusted models for specific countries.

**Keywords** Mortality · Annuity pricing · Lee–Carter model · Outlier-adjusted

## 1 Introduction

Life expectancy has increased significantly over the last century due to improvements in medicine, technology, and awareness in health issues, especially for developed countries. For instance, from 1960 to 2010, the life expectancy at birth in Canada, Japan, and the United Kingdom (UK) are increased by 15%, 18%, and 13%, respec-

---

C. Yavrum · A. S. Selcuk-Kestel (✉)

Institute of Applied Mathematics, Actuarial Sciences, Middle East Technical University, Ankara, Turkey

e-mail: [skestel@metu.edu.tr](mailto:skestel@metu.edu.tr)

C. Yavrum

e-mail: [cyavrum@metu.edu.tr](mailto:cyavrum@metu.edu.tr)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

495

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_30](https://doi.org/10.1007/978-3-030-85254-2_30)

tively, whereas, in Bulgaria, this change is around 8.9% (Human Mortality Database 2021). This variation is remarkably low in Russia whose life expectancy at birth shows an increase of only 0.3% (from 68.70 to 68.92) (Human Mortality Database 2021). These differences between countries can be explained by the causes such as environmental conditions, economic crisis, low incomes, the weak healthcare system and some extraordinary events that are outbreak of war, occurrence of pandemics, and radical changes in economics or politics (Chang et al. 2016). As a consequence, mortality rates which are the main component of pricing annuities, life insurance, retirement payments may result in misleading predictions and evaluations to set up a decision-making, like critical policy decisions on the retirement and insurance systems. The insurance premiums and annuities, one of the most significant income items of the financial sector, are calculated using the estimated future mortality rates. Therefore, using future mortality estimates generated by the model that does not include all of these factors can impact the financial strength of the life insurance industry and the stability of the pension system of a country.

In addition to its variation by time and age, mortality trends may have outliers (jumps) at specific time dimensions and age points. Mortality jumps are rare, but their presence could alter the long-term mortality trends by triggering many unexpected deaths, thereby affecting future estimates. As Stracke and Heinen (2021) state, additional claims received from unexpected pandemic would cost nearly €5 billion (50% of the market's total annual gross profit) for the German insurance market, as not even being the worst scenario. Another example, the earthquake and tsunami that occurred in southern Asia in 2004 made nearly 130,000 people missing and killed 180,000 (Carpenter 2005). If such an event would occur in a more economically developed area, the life insurance industry would deal with handling catastrophic losses (Carpenter 2005).

In light of all these significant indicators, companies should model the dynamics of mortality over time and to achieve this, a number of stochastic mortality models are introduced in the literature. These models are divided into two of which the first one is paying attention to the force of mortality using continuous time processes, and the second one concentrates mortality rates directly by discrete-time processes. While some continuous time models are pointed out by Biffis (2005) and Cairns and Dowd (2006), Lee and Carter (1992) developed the most famous one originated by a new method for extrapolation of age patterns and trends in mortality. Despite some drawbacks, its simplicity and quick-applicability make the Lee–Carter model to be utilized by some census bureaus, such as Census Bureau of the United States (Hollmann et al. 2000).

Regardless of the mortality estimation method, the presence of mortality jumps could lead to influences both in the sample and partial autocorrelation functions, which may cause erroneous predictions in the model (Li and Chan 2007). The use of the inadequate model affects pricing and reserve allocation for life insurances and annuities. This situation poses big threats to the solvency and price competitiveness of life insurance companies. Hence, appropriate mortality forecasting models should be used to predict this situation (Brouhns et al. 2002). Lee–Carter model has been extended by Renshaw and Haberman (2003) and Brouhns et al. (2002), where

Renshaw and Haberman describe a method for modeling reduction factors using regression methods within the generalized linear modeling framework, Brouhns et al. use Poisson regression model to forecast age–sex-specific mortality rates. However, none of these approaches include possible random outliers in the mortality data. Lin and Cox (2008), Cox et al. (2006), Chen and Cox (2009) improve their models to allocate outlier locations using a discrete-time Markov chain, Poisson distribution and independent Bernoulli distribution, respectively. Li and Chan (2005) use an approach to take into account possible outliers based on the Lee–Carter model. They create outlier-adjusted time series used with Lee–Carter model by using the iteration process developed by Chen and Liu (1993) to determine the locations of outliers and adjust their effects. Chang et al. (2016) develop the iteration process to incorporate outlier effects. After that, Chen and Liu (1993) improve their process for the joint estimation of model parameters and outlier effects. The securitization of insurance industry is critical as its strength reflects the economic stability of the country. The Global Economy (TheGlobalEconomy.com 2021) indicates that the average of insurance company assets including annuities and pension plans in the world as part of GDP is 16.48% in 2016. It can be easily said that a decrease in this number would have an impact on country's economy on its own. For this reason, the mortality jumps influencing the prices of annuities and pension liabilities also become a consideration of economic robustness.

There are studies that include the above-mentioned approaches to determine the effect of models incorporate with mortality jumps on securitization (Chen and Cox 2009; Liu and Li 2015). However, the effect of mortality jumps on the prices of life annuities and their liability are not covered depth in the literature. Due to the influence of sudden jumps in mortality, we aim to capture the outliers and we employ Outlier-Adjusted Lee–Carter model (Chen and Liu 1993; Li and Chan 2007) to the mortality rates of selected countries whose history exposes epidemics, wars, civil turbulences. To depict the influence of economic welfare, we select Canada, the UK, Japan as developed and old-age-dominated populations, whereas Bulgaria and Russia as emerging countries. To investigate the impact of mortality jumps on annuity price fluctuations, we make comparisons based on two approaches: (i) Outlier-Adjusted Lee–Carter (OALC) model and (ii) Lee–Carter (LC) model. We expect these evaluations illustrate the striking impact of mortality jumps on insurance pricing. The findings illuminate that the mortality jump model should be considered, especially for the countries that have disadvantages in the field of migration, economy, individual incomes, healthcare system, and/or underwent many wars and diseases in its past.

The chapter is organized as follows. Section 2 shows the methodology of the models and performance criteria as well as the iteration cycle that needs to be applied during the process. Section 3 contains the implementation, results of the applied models and the difference of annuity prices between selected countries. Section 4 concludes the results with a brief discussion.



## 2 Outlier-Adjusted Lee–Carter Model

Lee–Carter model which defines long-term mortality forecasts based on a combination of standard time series and an approach to handle the age distribution of mortality, is given as

$$\ln(m_{x,t}) = a_x + b_x \kappa_t + \epsilon_{x,t} \quad (1)$$

where  $m_{x,t}$  is the age-specific central death rate for age  $x$  at time  $t$ ,  $a_x$  stands for the age pattern of death rates,  $b_x$  is the age-specific reactions to the time-varying factor,  $\kappa_t$  indicates the mortality index in the year  $t$ , while  $\epsilon_{x,t}$  is the error term that captures the age-specific influences not reflected in the model for age  $x$  and time  $t$ . It is a widely known fact that the model is overparameterized. To obtain a unique solution,

$$\sum_x b_x = 1 \quad \text{and} \quad \sum_t \kappa_t = 0 \quad (2)$$

whose implementation to the model enables the age pattern of death rates,  $a_x$ , becoming the average value of  $\ln(m_{x,t})$  over time. Two-stage estimation procedure obtains the unique solution. We then ensure that the number of deaths deduced from model and the actual number of deaths are equal to each other. Additionally, Box and Jenkin's approach is employed to generate an autoregressive integrated moving average (ARIMA) model for estimating the mortality index,  $\kappa_t$ .

### 2.1 Outlier Modeling and Adjustment

Lee–Carter model which is the base for implementation constitutes an outlier-adjusted model with the help of an iteration cycle. The outlier analysis consists of two issues at which the first one is the determination of the location of the outlier values that may exist in the mortality index and the second one is finding and adjusting the effects of outliers if any exists. For the first issue, the value of standardized statistics of outlier effects should be found in order to detect outliers (Chang et al. 2016). For the latter one, more complex approaches and processes should be applied to standardized statistics of outlier effects (Chen and Liu 1993). Furthermore, there are two types of problems encountered in the outlier detection and adjustment procedure (Chen and Liu 1993). These are (i) having an outlier in mortality data may cause an error in model selection, (ii) even the model is selected correctly, the effect of outliers can significantly influence the estimation of model parameters. Chen and Liu's approach partly solves the second problem, while the first one stays the same. Regardless of the above-mentioned shortcomings, we use this method in our analysis, as it is the newest one in the literature.

Let  $Z_t$  be an outlier-free time series following an ARIMA(p,d,q),

$$\phi(B)(1 - B)^d Z_t = \theta(B)\alpha_t \tag{3}$$

where  $B$  is the backshift operator such that  $B^s Z_t = Z_{t-s}$  and  $\alpha_t$  represents white noise random variable with mean zero and constant variance  $\sigma^2$ .

Time series with outliers can be formed as outlier-free time series plus the effects of emergent outliers,  $\Delta_t(T, w)$ , where  $T$  and  $w$  are the location and the size of an outlier, respectively. Then, the series becomes,

$$Y_t = Z_t + \Delta_t(T, w) \tag{4}$$

Four types of outliers (Tsay 1988; Chen and Liu 1993; Li and Chan 2005) are Innovational Outlier (IO), Additive Outlier (AO), Temporary Change (TC), and Level Shift (LS). While an AO influences only a single observation, an IO affects all observations after  $T$  year with some decreasing pattern until the effect vanishes. This situation is slightly different for TC and LS. The effect of outlier remains the same influencing all observations after  $T$  year for LS and it decreases until to reach zero point by linearly for TC. It is stated commonly that a large portion of the outliers comprises of AO and IO (Chang et al. 2016). Since we focus on short time effects of mortality jumps that arise from extraordinary events affecting mortality rates for a short time, we consider these two intervention models, AO and IO, which are expressed as

$$\Delta_{t_{Ao}}(T, w) = wD_t^T \tag{5}$$

$$\Delta_{t_{Io}}(T, w) = \frac{\theta(B)}{\phi(B)(1 - B)^d} wD_t^T \tag{6}$$

respectively. Here,  $D_t^T$  is a binary variable having 1 in the presence of outliers at time  $T$ .

The outlier detection method (Chang et al. 2016) is grounded on the effects of outliers on the residuals of the model. The values of standardized statistics of outlier effects should be calculated to detect possible outliers. To achieve this,  $Z_t$  is expressed with its polynomial function of  $\pi(B)$ ,

$$\pi(B) = \frac{\phi(B)(1 - B)^d}{\theta(B)} = 1 - \pi_1 B_1 - \pi_2 B_2 - \dots \tag{7}$$

where  $\pi_j$  weights of outliers found at location  $T$ , influencing the years after  $T$  so,  $j \geq T$ . While the distance between  $j$  and  $T$  increases,  $\pi_j$  becomes zero as the effect of an outlier at location  $T$  does not impact on distance mortality values.

Given

$$\pi(B)Z_t = \alpha_t, \tag{8}$$

the Eq. 4 becomes,

$$\hat{e}_t = \pi(B)Y_t \quad \text{for } t = 1, 2, 3, \dots \tag{9}$$

where  $\hat{e}_t$  defines the residuals from the time series with outliers. In terms of considered outlier types, AO and IO, the residuals become

$$\hat{e}_{t_{AO}} = wD_t^T + \alpha_t \tag{10}$$

$$\hat{e}_{t_{IO}} = w\pi(B)D_t^T + \alpha_t \tag{11}$$

Equations 10 and 11 can be symbolized as a general time-series structure as follow:

$$\hat{e}_t = wd(l, t) + \alpha_t \tag{12}$$

where  $l = (AO, IO)$ ;  $d(l, t) = 0$  with  $t < T$ ;  $d(l, T) = 1$  for both types. When  $k \geq 1$ ,

$$d(AO, T + k) = 0, \quad k = 1, 2, 3, \dots \tag{13}$$

$$d(IO, T + k) = -\pi_k, \quad k = 1, 2, 3, \dots \tag{14}$$

It is clear to reach the conclusion that the effect of an AO is contained only at a particular point  $T$ , whereas the effect of IO is dispersed through the time after at time point  $T$ . Consequently, from the least squares theory, the effect of an outlier at  $t = t_1$  can be formed as

$$\hat{w}_{AO}(t_i) = \hat{e}_{t_i}, \quad i = 1, 2, 3, \dots \tag{15}$$

$$\hat{w}_{IO}(t_i) = \frac{\sum_{t=t_1}^{t_n} \hat{e}_t d_{IO,t}}{\sum_{t=t_1}^{t_n} d_{IO,t}^2}, \quad i = 1, 2, 3, \dots \tag{16}$$

where  $t_n$  indicates the last attainable age.

For locating possible outliers, we analyze the maximum value of the standardized statistics (Chang et al. 2016) as,

$$\hat{\tau}_{AO}(t_i) = \frac{\hat{w}_{AO}(t_i)}{\hat{\sigma}_\alpha}, \quad i = 1, 2, 3, \dots, \tag{17}$$

$$\hat{\tau}_{IO}(t_i) = \frac{\hat{w}_{IO}(t_i)}{\hat{\sigma}_\alpha} \left( \sum_{t=t_1}^n d_{IO,t}^2 \right)^{1/2}, \quad i = 1, 2, 3, \dots \tag{18}$$

Here,  $\hat{\sigma}_\alpha$  is the estimate of residual standard deviation. Hence, the possible location of an outlier can be determined when the standardized values exceed a chosen constant value of  $C$ .

In order to decide whether the outlier is a form of AO or IO when both of their effects are greater than  $C$ , we follow a simple rule that chooses the type of outlier whose effect is greater than the others (Society 2010). To achieve a high degree of sensitivity in locating the outliers and to be consistent with the literature (Chang et al. 2016), we take the value  $C$  as 3.0.

As next, the standard deviation of residuals should be calculated to reach the numerical value for the maximum of standardized statistics given by the Eqs. 17 and 18. To calculate residual standard deviation, the median absolute deviation is used. Given  $\tilde{e}$  is the median of the estimated residuals,

$$\hat{\sigma}_\alpha = 1.483 \times \text{median}\{|\hat{e}_t - \tilde{e}|\} \quad (19)$$

is quantified to find the estimates of  $\hat{\tau}_{AO}(t_i)$  and  $\hat{\tau}_{IO}(t_i)$ ,  $i = 1, 2, 3, \dots$

Once the outlier locations are set, the outlier adjustment is needed to estimate new model parameters and hence the outlier effects. To accomplish this, an iteration cycle repeated until no more outliers are found, is needed. When the iteration stops, ultimate ARIMA and its parameters are used to forecast the mortality index,  $\kappa_t$ .

The iteration process whose algorithm is given in detail (Algorithm 1) starts with determining the order of the underlying ARIMA for  $\kappa_t$  in LC model. After this step, the residuals are needed as the next stage. Then, the coefficient of  $\pi(B)$ , outlier effects of AO and IO, and standard deviation of the residuals are evaluated. For determining whether there is an outlier in the series, the absolute value of standardized statistics for all time points are computed and compared with the pre-determined  $C$  value. This pre-determined value can be adjusted according to the strength of the model to locate the jumps. If none of the absolute values of the standardized statistics exceeds  $C$  value, then the series is taken as outlier-free or outlier-adjusted. Otherwise, the effects of outliers should be removed from the residuals. With the new residuals, the standard deviation is found again for the possibility of detecting new outliers. This process continues until no further outlier is found. The final ARIMA and its parameters are used to create an outlier-adjusted Lee–Carter model for the ultimate mortality index.

During the iteration process, the number of parameters and the variance of the model may change. Thus, they can have an impact on the accuracy of models. We measure the performance of the models using the Akaike Information Criterion (AIC). The ultimate ARIMA and its parameters are employed to forecast the mortality index. Afterward, the death probability of age  $x$ ,  $\hat{q}_x$ , is calculated as

$$\hat{q}_x = \frac{\hat{m}_x}{1 + (1 - c_x)\hat{m}_x} \quad (20)$$

Here,  $\hat{m}_x$  is the forecasted central death rate for age  $x$ ,  $c_x$  is the average number of years lived within the age interval  $x$  and  $x + 1$ . As in the Human Mortality Database protocol (Human Mortality Database 2021),  $c_x$  is taken as 0.5 for all ages except zero. For the beginning age, the last observed numbers in the data for all ages are used.

---

**Algorithm 1:** Iteration Process
 

---

- Step 1:* Use Box and Jenkin's approach to identify the orders  $p, d, q$ .
- Step 2:* Compute the residuals of mortality index found from LC model.
- Step 3:* Calculate the coefficient of  $\pi(B)$  and then, outlier effects of AO and IO accordingly.
- Step 4:* Evaluate the standard deviation of residuals obtained in *Step 2*.
- Step 5:* Compute standardized statistics for AO and IO for all time points, decide whether there is an outlier. Then, determine the type of outliers by comparing values with  $C$ .
- Step 6:* If no outlier is found in *Step 5*, then Stop: Assign the series is outlier-free or outlier-adjusted. Otherwise, remove the effects of outliers by defining new residuals for AO and IO.
- Step 7:* Re-calculate the standard deviation of residuals with adjusted residuals and go to *Step 5*. Repeat this cycle until no further outlier can be identified.
- Step 8:* After the locations of all possible outliers are found, remove the effects of outliers from the mortality index to calculate new mortality index.
- Step 9:* Go to *Step 1* using new series based on the mortality index and repeat the cycle until no further outlier is found.
- Step 10:* The final ARIMA and its parameters are used to create the ultimate forecasting model for mortality index.
- 

### 3 Implementation of the Models

Even though human life is restricted and life expectancy has more or less a similar pattern, country, geographic location, economic development, climate, race, and political–social effects cause variations. The time influence is also an important factor in aging structure. We choose five countries of those two are located in Central Europe, one in North America, one in Central Asia, and one is in Far East. Along with their geographical differences, the UK and Japan do have longevity issues, though their developed economies. Nevertheless, Russia and Bulgaria are accounted as emerging countries. Except for Canada, all countries experience big influence of wars and epidemics. Russia and Bulgaria are mostly prone to political risks due to reformist changes in the twentieth century. In terms of social welfare, the UK, Canada, and Japan have long history in life and pension insurance and annuity products, whereas Russian and Bulgarian insurance markets are flourishing.

Mortality data is collected using Human Mortality Database (2021) with single age based (0–110). The summary of the data and related parameters are given in Table 1. Among the selected countries, the longest period belongs to Canada with 96 years of mortality rates, whereas the Russian rates goes back to only 56 years. The forecasts are projected till 2060.

**Table 1** Specific properties of mortality data

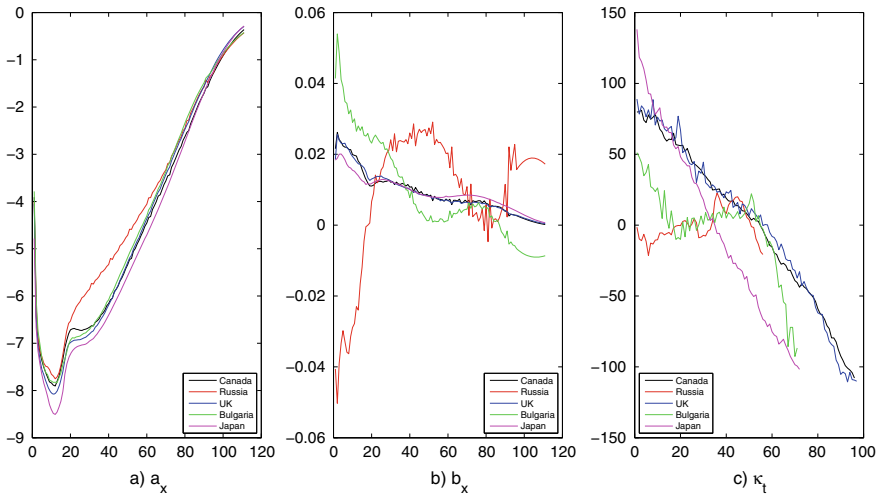
			Annuity	
	Country	Years	Ages	Periods
Developed	Canada	1921–2016		
	The UK	1922–2016	0	5 Yrs
	Japan	1947–2017	30	10 Yrs
Emerging	Russia	1959–2014	70	30 Yrs
	Bulgaria	1947–2010		

### 3.1 Model Estimations

Using the steps explained in previous sections, we estimate the parameters of LC model for each country. The plots of estimated values of  $a_x$ ,  $b_x$ , and  $\kappa_t$  are represented in Fig. 1.  $a_x$ , the age pattern of death rates;  $b_x$ , age-specific reaction time factor;  $\kappa_t$ , mortality index. Although there is no significant difference between countries based on age pattern of death rates, great differences are seen for  $b_x$  and  $\kappa_t$ . The plot of  $a_x$  shows that the average values of  $\ln(m_{x,t})$  over the years are quite similar between countries. However, every age reacts to mortality improvements differently, as seen in  $b_x$  plot, meaning that while mortality improvements benefit younger generation in Canada, they provide more benefit to adults in Russia. This pattern can be seen for Bulgarian data too. Bulgarian population reacts unequally to mortality improvements as the mortality index of younger age groups benefit much more than older age groups. The plot of  $\kappa_t$  clarifies that mortality rates decrease significantly in Canada, Japan, and the UK thus, create problems on longevity risk; Russia and Bulgaria have fluctuating mortality trends creating unstable mortality rates over the years.

The locations, types, and effects of outliers are detected using the iteration cycle for Russia and Canada. In Table 2, along with orders of ARIMA and its required parameters, the estimates of  $c_o$  and  $\hat{\sigma}^2$  representing the constant value within the model and the estimation of the variation of  $\kappa_t$ , respectively, are summarized. The effect of outlier  $w$  and its standardized value ( $\tau$ ) are listed. When no outliers are detected consequently two times, the iteration process stops and enables us to reach outlier-adjusted mortality index.

For Russian data, the iteration starts with ARIMA(3,1,2) and reaches ARIMA(0,1,0) after 23 iterations. During those iterations, many outliers from both types are identified and their effects are removed from residuals and mortality index simultaneously. The type of outlier identified in the last observation of time-series cannot be distinguished between AO and IO, as emphasised by Chen and Liu (1993), as seen in Russian data. Another significant observation deduced from the table is that outlier-adjusted models have a smaller  $\hat{\sigma}^2$  than starting models indicating that the outlier-adjusted models are superior to original models. A larger variance for  $\kappa_t$  affects the performance of the model, so it should be reduced in order to capture a better model. For Canadian data, the orders of ARIMA do not change even after 18



**Fig. 1** Lee–Carter model parameter estimates

iterations but removing the effects of outliers help us have a smaller  $\hat{\sigma}^2$  for the model. The same analyses on the UK, Japan, and Bulgaria are performed and presented in Table 3.

We determine that compared to Russia and Bulgaria, fewer outliers are found for the UK and Japan. For Japan, only one outlier with AO type is identified in time point two. After this effect is removed, the outlier-adjusted mortality index follows ARIMA(2,1,2). For the UK, the orders of ARIMA do not change where the variance of  $\kappa_t$  decreases significantly from 25.20 to 17.50 in four iterations. More drastic changes are observed for Bulgarian data. Six outliers which all of them are AO type, are found in nine iterations. Outlier-adjusted mortality index is constructed based on ARIMA(0,1,3) even though iteration starts with ARIMA(0,1,0). Although the ultimate models do not change for the UK and Canada, all countries depict variation so that the models generated from Lee–Carter do not reflect the historical data truly.

Although not all outliers identified in the iteration process, some of them may be associated with actual historical events. When the corresponding years of outliers are analyzed, it can be said that wars and economic crises play a major role in changing mortality rates thus, having outliers in mortality data. The Second World War affects the UK in 1940 and 1942, which is associated with mortality outliers. In addition to wars, some important accidents such as Chernobyl in 1986 in Ukraine (former USSR) influence more than one country increasing their mortality rates. After the dissolution of the USSR in 1990 causing separation into 15 different countries makes fluctuations in mortality rates between 1993 and 1996 for Russia as well. Moreover, the effects of economic crises can be seen in Canada between 1929 and 1939 identified as the Great Depression, in Russia between 2011 and 2014 having a sharp decrease in income and purchasing power that may be explanatory of increase in mortality

**Table 2** Outlier detection for Russian and Canadian mortality

Country	ARIMA	Iteration	Parameters					Outliers					
			$c_0$	$\theta_1$	$\theta_2$	$\theta_3$	$\phi_1$	$\phi_2$	$\hat{\sigma}^2$	Time	Type	$w$	$\tau$
Russia	(3,1,2)	1	-0.93	-0.01	-0.75	0.28	0.33	1.00	13.04	6	IO	-5.34	-3.10
		2	-0.09	-0.10	0.70	0.01	0.18	-0.63	18.83	6	IO	-5.95	-4.28
	(0,1,0)	3	-0.04	1.54	-0.94	0.22	-1.31	0.56	16.42	7	AO	14.27	6.27
		4	-0.52	0.14	-0.67	-0.05	-0.03	0.84	18.01	53	AO	-6.84	-3.01
	(0,1,0)	5	-0.33	-	-	-	-	-	20.50	7	AO	7.04	3.00
		6	-0.33	-	-	-	-	-	18.26	36	AO	17.65	4.23
	(0,1,0)	7	-0.33	-	-	-	-	-	18.26	37	AO	14.10	3.55
		8	-0.33	-	-	-	-	-	15.79	7	IO	-8.43	-3.09
	(0,1,0)	9	-0.33	-	-	-	-	-	14.76	7	IO	-8.38	-3.21
		10	-0.33	-	-	-	-	-	20.09	8	AO	14.97	4.07
	(0,1,0)	11	-0.33	-	-	-	-	-	20.09	-	-	-	-
		12	-0.33	-	-	-	-	-	23.16	7	IO	11.75	4.20
	(0,1,0)	13	-0.33	-	-	-	-	-	18.65	8	IO	-12.40	-4.43
		14	-0.33	-	-	-	-	-	29.45	8	IO	-18.53	-6.55
									9	AO	20.88	5.20	
									8	IO	-8.27	-3.07	
									9	AO	18.65	4.88	
									-	-	-	-	

(continued)



Table 2 (continued)

Country	ARIMA	Iteration	Parameters					Outliers					
			$c_0$	$\theta_1$	$\theta_2$	$\theta_3$	$\phi_1$	$\phi_2$	$\hat{\sigma}^2$	Time	Type	$w$	$\tau$
Canada	(2,1,2)	15	-0.25	0.38	-0.56	-	-0.86	1.00	20.78	8	AO	17.73	4.27
		16	-0.06	0.24	0.47	-	-0.07	-0.47	13.57	-	-	-	-
	(0,1,0)	17	-0.33	-	-	-	-	-	14.49	38	AO	9.20	3.03
		18	-0.33	-	-	-	-	-	15.31	54	AO	-10.33	-3.40
	(0,1,0)	19	-0.33	-	-	-	-	-	15.31	54	IO	7.68	3.77
		20	-0.33	-	-	-	-	-	13.87	55	AO	-15.30	-5.38
	(0,1,0)	21	-0.33	-	-	-	-	-	13.87	56	AO/IO	-9.86	-3.39
		22	-0.17	-	-	-	-	-	12.10	-	-	-	-
(0,1,0)	23	-0.17	-	-	-	-	-	12.10	-	-	-	-	
(0,1,0)	1	-1.90	-	-	-	-	-	5.38	6	AO	7.32	3.13	
	2	-1.90	-	-	-	-	-	5.34	17	IO	6.10	3.73	
	3	-1.90	-	-	-	-	-	6.47	17	IO	6.27	4.09	
	4	-1.90	-	-	-	-	-	6.47	18	AO	-12.48	-5.66	
	5	-1.90	-	-	-	-	-	4.49	-	-	-	-	
	6	-1.90	-	-	-	-	-	5.77	17	IO	-9.07	-5.87	
	7	-1.90	-	-	-	-	-	7.81	18	AO	12.37	5.58	
	8	-1.90	-	-	-	-	-	9.80	18	AO	6.92	4.35	
									18	IO	6.46	4.30	
									19	AO	-11.70	-4.95	
									19	AO	-7.26	-3.09	
									-	-	-	-	

(continued)

**Table 2** (continued)

Country	ARIMA	Iteration	Parameters						Outliers					
			$c_0$	$\theta_1$	$\theta_2$	$\theta_3$	$\phi_1$	$\phi_2$	$\hat{\sigma}^2$	Time	Type	$w$	$\tau$	
	(1,1,0)	9	-2.49	-0.35	-	-	-	-	8.60	18	IO	-13.58	-7.57	
		10	-1.79	0.06	-	-	-	-	4.71	19	AO	12.30	5.40	
			-	-	-	-	-	-	-	4.26	19	AO	9.27	4.26
			-	-	-	-	-	-	-	3.43	20	AO	7.47	3.43
			-1.91	-0.01	-	-	-	-	5.98	9	IO	5.02	3.07	
			-1.97	-0.04	-	-	-	-	5.75	11	AO	-6.62	-3.20	
	-2.00	-0.06	-	-	-	-	6.13	-	-	-	-			
	(0,1,0)	14	-1.90	-	-	-	-	-	6.15	19	IO	-7.62	-4.99	
		15	-	-	-	-	-	-	-	6.69	20	AO	6.69	3.04
			-1.90	-	-	-	-	-	4.76	11	IO	4.50	3.05	
			-	-	-	-	-	-	-	-	19	IO	-4.55	-3.08
			-	-	-	-	-	-	-	-	20	AO	7.83	3.69
	16	-1.90	-	-	-	-	-	5.56	11	IO	4.64	3.25		
17	-1.90	-	-	-	-	-	5.67	12	AO	-9.58	-4.66			
<b>(0,1,0)</b>	<b>18</b>	<b>-0.32</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>4.13</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>		

Table 3 Outlier detection for Briton, Japanese, and Bulgarian mortality

Country	ARIMA	Parameters						Outliers						
		Iteration	$c_0$	$\theta_1$	$\theta_2$	$\theta_3$	$\phi_1$	$\phi_2$	$\phi_3$	$\hat{\sigma}^2$	Time	Type	$w$	$\tau$
The UK	(0,1,1)	1	-2.09	-	-	-	-0.51	-	-	25.20	8	IO	13.28	3.16
		2	-2.09	-	-	-	-0.48	-	-	18.91	19	IO	17.14	4.08
		3	-2.09	-	-	-	-0.44	-	-	17.50	9	AO	-14.34	-3.01
Japan	(0,1,1)	4	-2.09	-	-	-	-0.44	-	-	17.50	21	AO	-18.14	-3.81
		1	-0.47	0.84	-	-	-1.24	0.16	0.08	9.67	2	AO	-11.61	-3.41
		2	-2.44	0.30	-	-	-0.47	-0.02	0.29	11.39	-	-	-	-
Bulgarian	(2,1,2)	3	-1.46	1.20	-0.62	-	-1.53	1.00	-	8.46	-	-	-	-
		1	-2.27	-	-	-	-	-	-	108.61	63	AO	-41.04	-4.10
		2	-2.27	-	-	-	-	-	-	102.30	-	-	-	-
	(1,1,2)	3	-1.29	0.58	-	-	-0.94	0.61	-	80.06	64	AO/IO	-26.97	-3.19
		4	-0.21	0.95	-	-	-1.37	0.55	-	64.44	53	AO	-24.11	-3.00
		5	-1.52	0.33	-	-	-0.79	0.68	-	66.49	21	AO	22.47	3.01
	(3,1,0)	6	-1.71	0.27	-	-	-0.65	0.69	-	66.13	62	AO	-23.68	-3.18
		7	-1.49	-0.19	0.04	0.51	-	-	-	59.24	23	AO	21.43	3.02
		8	-1.51	-0.17	0.05	0.48	-	-	-	59.11	-	-	-	-
	(0,1,3)	9	-2.08	-	-	-	-0.14	0.21	58.31	-	-	-	-	

**Table 4** Model performance indicators for selected countries

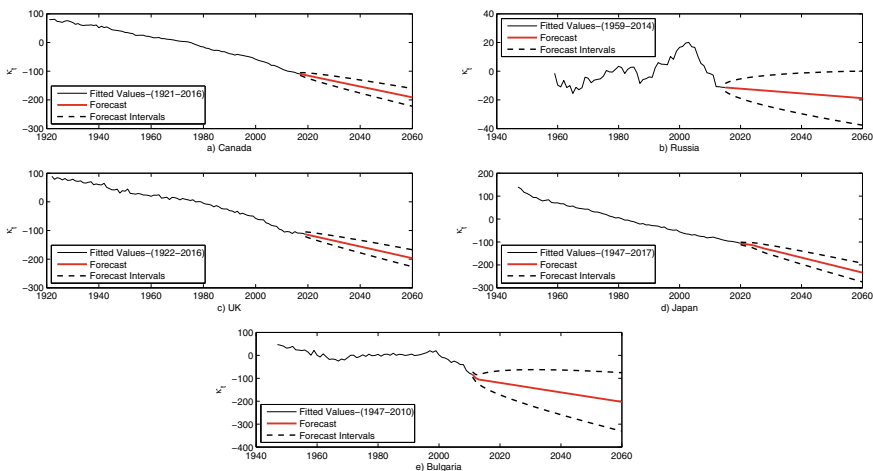
	Values of AIC	
	LC	OALC
Canada	653.97	477.68
Russia	130.05	120.02
The UK	306.44	285.55
Japan	186.12	163.64
Bulgaria	298.76	280.17

rates. It is also interesting to note that Canada does not contain any outliers after 1940, whereas Russia has outliers identified in the twenty-first century explaining continuing unstable mortality trends.

The comparison of solely implementation of LC with OALC model is made using AIC statistics and the results are given in Table 4. The smaller AIC is achieved when the outlier influence is incorporated into LC model.

### 3.2 Forecasting the Mortality Rates

The mortality index is forecasted until 2060 for all countries using outlier-adjusted ARIMA with 95% confidence interval and presented in Fig. 2. Since  $\kappa_t$  for Russian and Bulgarian data have more fluctuations in their mortality rates, the confidence intervals pose wider bands. As expected for the UK, Japan, and Canada, the forecasted



**Fig. 2** Forecasted mortality index for selected countries

mortality index decreases smoothly till 2060, whereas we may observe some increase for Russia and Bulgaria as their confidence intervals cover an increasing part in the mortality index as well. Using the forecasted mortality index, OALC model enables us to estimate the survival probabilities. Thus, the expected life expectancy at birth can also be constructed based on survival probabilities.

### 3.3 *Annuity Pricing and Portfolio Impacts*

Life annuities are financial products sold by insurance companies paid annually or at different intervals beginning at a stated year. Annuities are generally purchased by investors who aim to provide a fixed income after their retirement. There are two perspectives in terms of annuities: (i) buyers that make regular payments to the insurance company in the period called accumulation and (ii) companies make payments to buyers. In addition, both parties may pay the sum of the annuity in advance so, subsequent payments are calculated accordingly. It is known that even slight fluctuation in price can affect both parties deeply. For the company which is responsible for hundreds of thousands of annuitant payments, a small change in annuity price can cause a huge deficit in its financial budget.

The prices of whole life annuities evaluated at the end of the year (due) for the specific ages 0, 30, and 70 in 2060 for both LC and OALC models are presented in Table 5. Here, the interest rate  $i$  is chosen an arbitrary value of 3%. It can be depicted that the life annuity prices are reduced with respect to age for both models and all countries. As longevity is an important consideration, Japan has the highest prices for all ages in both models followed by Canada and the UK. It is clear to notice that Russia has the lowest annuity prices for all ages compared to other countries, also indicating that Russia has high mortality rates. To illustrate the performance of OALC model, the price differences between LC and OALC are presented. The difference values are positive for all ages in Canada and Japan, whereas take negative values for the UK and Bulgaria. For Russia, the difference is positive for beginning age though it is negative for the age of 30 and 70, representing that Russian mortality contains serious fluctuations over the years. The maximum absolute change is achieved for Russia, followed by Japan.

For a more comprehensive assessment, we create a portfolio consist of 10,000 insureds randomly generated from a uniform distribution between ages 15 and 75. Then, the portfolio is used for four different annuity types: (i) 5 year term, (ii) 10 year term, (iii) 30 year term, and (iv) Whole life. The annuity prices are calculated for both LC and OALC models and represented as the sum of the portfolio given in Table 6. The price increases as the number of term increases. While the price difference of whole life between models for Russia is around 5%, this difference does not exceed 1% for other countries. For the UK, the performances of models are so close to each other as there are almost no differences in four annuity types. In larger and

**Table 5** Prices of whole life annuities (due) in 2060

	Age	Whole life annuity price		
		LC	OALC	Difference (%)
Canada	0	31.59	31.66	0.22
	30	27.57	27.73	0.56
	70	14.48	14.77	2.01
The UK	0	31.53	31.52	−0.01
	30	27.38	27.38	−0.01
	70	13.95	13.95	−0.01
Japan	0	31.97	32.04	0.22
	30	28.46	29.39	5.02
	70	16.36	16.73	2.25
Russia	0	25.53	28.43	11.36
	30	25.36	24.13	−4.82
	70	10.08	9.59	−4.87
Bulgaria	0	30.84	30.81	−0.10
	30	25.60	25.53	−0.28
	70	10.44	10.37	−0.71

**Table 6** Prices of life annuities (due) in 2060 for portfolio

Annuity type		Life annuity price				
		Canada	Russia	The UK	Japan	Bulgaria
5 year term	LC	55,461	54,750	54,190	55,641	50,500
	OALC	55,495	54,366	54,190	55,665	50,015
	Difference (%)	0.06	−0.70	0.00	0.04	−0.96
10 year term	LC	93,980	91,298	93,808	94,668	92,397
	OALC	94,111	90,061	93,806	94,762	92,269
	Difference (%)	0.14	−1.35	0.00	0.10	−0.14
30 year term	LC	190,730	174,350	189,280	196,450	177,210
	OALC	191,730	168,310	189,260	197,410	176,690
	Difference (%)	0.52	−3.46	−0.01	0.49	−0.29
Whole life	LC	235,410	205,380	232,350	248,130	208,430
	OALC	237,500	194,990	232,320	250,480	207,610
	Difference (%)	0.89	−5.06	−0.01	0.95	−0.39

more diverse portfolios, the difference between prices can be more pronounced. Nevertheless, Canadian, Briton, and Japanese mortality data can be considered to be less affected by the model modification.

## 4 Conclusion

The effect of possible outliers in the mortality data on the price of life annuities may vary according to race, geographic location, economic welfare, and demographic structures. For this reason, Canadian, Briton, and Japanese mortality representing developed countries, Russia and Bulgaria as emerging markets are selected. By such a comparison, we aim to reach a better explication of the differences in mortality behaviors between countries. Moreover, we also compare in terms of four different annuity types to depict the product influence. In order to check the accuracy and sensitivity on a portfolio, calculations are also made on a simulated group of 10,000 insureds whose ages are randomly assigned between 15 and 75.

This study exposes that using the outlier-adjusted model in forecasting mortality rates is critical to the annuity prices for the countries with outliers in their mortality rates. It is observed that Russia is the most affected country as the annuity price differences in all scenarios come up to be the largest compared to others. On the other hand, the differences do not exceed 1% for the UK, Japan, and Canada as developed countries facing longevity risk for their populations. It is shown that as an emerging market, Bulgaria does not show sensitivity to outliers. The political reforms in Russia can be taken as a consequence of sensitivity to outliers.

## References

- Biffis E (2005) Affine processes for dynamic mortality and actuarial valuations. *Insur Math Econ* 37:443–468. <https://doi.org/10.1016/j.insmatheco.2005.05.003>
- Brouhns N, Denuit M, Vermunt JK (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insur Math Econ* 31:373–393. [https://doi.org/10.1016/S0167-6687\(02\)00185-3](https://doi.org/10.1016/S0167-6687(02)00185-3)
- Cairns AJG, Dowd K (2006) A two-factor model for stochastic mortality with parameter uncertainty 73:687–718
- Carpenter G (2005) Tsunami: Indian Ocean event and investigation into potential global risks. See the report release in Mar 2005
- Chang I, Tiao GC, Chen C, Chang I (2016) American Society for quality estimation of time series parameters in the presence of outliers linked references are available on JSTOR for this article: estimation of time series parameters in the presence of outliers 30:193–204
- Chen C, Liu L-M (1993) Joint estimation of model parameters and outlier effects in time series. *J Am Stat Assoc* 88:284. <https://doi.org/10.2307/2290724>
- Chen H, Cox SH (2009) Modeling mortality with jumps: applications to mortality securitization. *J Risk Insur* 76:727–751. <https://doi.org/10.1111/j.1539-6975.2009.01313.x>
- Cox SH, Lin Y, Wang S (2006) Multivariate exponential tilting and pricing implications for mortality securitization. *J Risk Insur* 73:719–736. <https://doi.org/10.1111/j.1539-6975.2006.00196.x>

- Fox AJ (2010) Outliers in time series. *J R Stat Soc Ser B (Methodol)* 34(3):350–363 (Blackwell Publishing for the Royal Statistical Society). <http://www.jstor.org/stable>
- Hollmann F, Mulder T, Kallan JE (2000) Methodology and assumptions for the population projections of the United States, 1999–2100. *Popul Div Work Pap* 38:1–24
- Human Mortality Database (2021) <https://www.mortality.org>. Accessed 22 Mar 2021
- Lee RD, Carter LR (1992) Modeling and forecasting U. S. mortality. *J Am Stat Assoc* 87:659. <https://doi.org/10.2307/2290201>
- Li SH, Chan WS (2005) Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. *Scand Actuar J* 2005:187–211. <https://doi.org/10.1080/03461230510006973>
- Li SH, Chan WS (2007) The Lee-Carter model for forecasting mortality, revisited. *N Am Actuar J* 11:68–89. <https://doi.org/10.1080/10920277.2007.10597438>
- Lin Y, Cox SH (2008) Securitization of catastrophe mortality risks. *Insur Math Econ* 42:628–637. <https://doi.org/10.1016/j.insmatheco.2007.06.005>
- Liu Y, Li JSH (2015) The age pattern of transitory mortality jumps and its impact on the pricing of catastrophic mortality bonds. *Insur Math Econ* 64:135–150. <https://doi.org/10.1016/j.insmatheco.2015.05.005>
- Renshaw A, Haberman S (2003) Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections. *J R Stat Soc Ser C Appl Stat* 52:119–137. <https://doi.org/10.1111/1467-9876.00393>
- Stracke A, Heinen W (2021) SOA—influenza pandemic: the impact on an insured lives life insurance portfolio. <https://www.soa.org/library/newsletters/the-actuary-magazine/2006/june/pub-influenza-the-impact-on-an-insured-lives-life-insurance-portfolio>. Accessed 22 Mar 2021
- TheGlobalEconomy.com (2021) <https://www.theglobaleconomy.com>. Accessed 22 Mar 2021
- Tsay RS (1988) Outliers, level shifts, and variance changes in time series. *J Forecast* 7:1–20. <https://doi.org/10.1002/for.3980070102>



# Optimal Life Insurance and Annuity Demand with Jump Diffusion and Regime Switching



Jinhui Zhang, Sachi Purcal, and Jiaqin Wei

**Abstract** Classic Merton optimal life-cycle portfolio and consumption models are based on diffusion models for risky assets. In this paper, we extend the Richard's (1975) optimal life-cycle model by allowing jumps and regime switching in the diffusion of risky assets. We develop a system of paired Hamilton–Jacobi–Bellman (HJB) equations. Using numerical methods, we obtain the results of agents' behaviour. Our findings are that agents would be more conservative in consumption and annuitisation when the economic environment is more volatile and the bequest motive is stronger. However, under certain conditions, agents might increase their exposure to risky assets.

**Keywords** Stochastic optimal control · Richard's model · Optimal investment · Jumps · Regime switching

## 1 Introduction

In this paper, we extend Richard's model (Richard 1975) to study investors' behaviour by allowing jumps and regime switching in the underlying asset dynamics. Based on empirical evidence that the underlying asset dynamics are impacted by change in the economic state, our motivation is the assumption that underlying asset dynamics could experience a sudden change due to a structural break in the economy which could be attributed to jumps and regime switching.

The presence of regime switching and jumps in the underlying asset dynamics is widely studied in the context of option pricing. A discontinuous model is proposed and examined in Merton (1976) for pricing options with an assumed

---

J. Zhang (✉)

Department of Actuarial Studies and Business Analytics, Macquarie Business School,  
Macquarie University, Sydney, Australia  
e-mail: [colin.zhang@mq.edu.au](mailto:colin.zhang@mq.edu.au)

S. Purcal · J. Wei

Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE,  
School of Statistics, East China Normal University, Shanghai, China

log-normal distributed jump size. Cont and Tankov (2004) described jumps in the underlying asset dynamics for option pricing via an exponential Lévy process model. Elliott et al. (2007) utilised a Markov-modulated pure jump process to derive the regime-switching HJB equation for European options, barrier options and American options. In Hamilton (1989), a discrete-state Markov process was introduced for regime-switching parameter values to allow endogenous structural breaks. Following Hamilton (1989), various numerical methods were presented in the literature for pricing options with regime switching. The lattice-based method, quadratic approximation method and front-fixing method were presented and studied in Bollen (1998), Brown (2001) and Wu and Kwok (1997), respectively.

Optimal investment strategy has been studied in the existing literature. The classic Merton model was developed in Merton (1969) with the assumed constant relative risk-aversion (CRRA) utility function. Richard (1975) generalised the Merton model by including the bequest motive and insurance demand in the model. To capture market dynamics, jumps and regime switching have been studied in the literature for optimal investment strategy. Based on Richard (1975), Purcal and Wang (2005) introduced a jump-diffusion environment with a fixed jump size. Hanson (2007) utilised CRRA utility and log-uniform jump amplitude to present optimal portfolio and consumption policies. Song et al. (2006) developed a numerical scheme for controlled regime-switching jump diffusions. For financial markets with regime switching, Zhang and Guo (2004) introduced near-optimal strategies and Sotomayor and Cadenillas (2009) presented explicit solutions for the optimisation problem of consumption and investment.

As they are made on a daily basis, financial decisions are evidently impacted by the changing market. According to Heaton and Lucas (2000), investor behaviours would deviate due to changes in the market risk. Shocks such as the Global Financial Crisis are documented as having an influence on the portfolio choices of investors (Bateman et al. 2011). Therefore, investors, especially retirees, face risks not only from consumption, investment and longevity but also from rapid changes in the market or economic state. However, compared to a large amount of research on the investment optimisation problem, the post-retirement optimal financial strategy problem has not received much attention (Gupta and Murray 2003). This motivates us to build a post-retirement model to study the retiree behaviours of consumption, investment and bequests when the state of the financial market state is volatile.

With inspiration from the existing literature, in this paper, we extend Richard's model to include jumps and regime switching in the financial market. Numerical results are obtained for the optimal consumption, investment and insurance strategies in order to study the bequest motive and market risk effects.

This paper is organised as follows. Section 2 extends the Richard's model to the regime-switching jump diffusion environment. Section 3.2 demonstrates the numerical results and analyses the findings while Sect. 4 concludes the paper.

## 2 Model and Method

From Hanson (2007), the dynamics of the risky asset price,  $X(t)$ , are assumed to be

$$dX(t) = X(t) (\alpha(t)dt + \sigma(t)dB(t) + Jd\psi(t)),$$

where  $\alpha(t)$  and  $\sigma(t)$  are the average return rate and volatility of the risky asset price, respectively,  $dB(t)$  is the standard Brownian motion,  $J$  is a uniform distributed jump amplitude on  $[\mathcal{G}_1(t), \mathcal{G}_2(t)]$  and  $\psi(t)$  is a discontinuous one-dimensional Poisson process with a jump rate  $\lambda$ .

The agent is assumed to have a random time of death which is modelled by the survival rate,  $S(t)$ , and the force of mortality,  $\mu(t)$ , with the density function of mortality,  $f(t) = \mu(t) \cdot S(t)$ .

Here we assume the agent has utility from consumption, that is,  $U_1$ , as well as a utility from leaving bequests, that is,  $U_2$ . Then the objective of a utility-maximising agent is

$$\max_{C(t), \pi(t), Z(t)} E_t \left[ \int_t^\tau \frac{S(T)}{S(t)} \left( \frac{\theta(T)}{\theta(t)} U_1(C, T) + \mu(T) \frac{\theta(T)}{\theta(t)} U_2(L, T) \right) dT \right],$$

which is subject to the dynamics of wealth  $W$ ,

$$dW = [(\alpha(t) - r(t))\pi(t)W + r \cdot W + Y - C(t) - P(t)]dt + \pi(t)\sigma(t)Wdq(t) + \pi W \sum_{k=1}^{d\psi(t)} J(T_k^-),$$

where  $r(t)$  is the risk-free rate, consumption,  $C(t)$ , the proportion of wealth invested in risky assets,  $\pi(t)$ , and legacy amount,  $L(t)$ , are the control variables,  $P(t)$  is the insurance premium,  $P(t) = \mu(t)(L(t) - W(t))$ ,  $Y$  is the deterministic income which is set to be zero for simplicity, and  $T_k^-$  is the pre-jump time.

A continuous-time Markov chain process  $\mathcal{X} := \{\mathcal{X}_t\}_{t \in \mathcal{T}}$  is defined here with a finite state-space  $\{e_1, e_2, \dots, e_N\}$ , where  $e_i = (0, \dots, 1, \dots, 0)' \in \mathcal{R}^N$ .

The element  $\mathcal{X}_t = e_i$  of the Markov chain demonstrates that, at time  $t$ , the economy is in the  $i$ th state.

Elliott et al. (1994) showed that the Markov chain process  $\mathcal{X} = \{\mathcal{X}_t, t \in \mathcal{T}\}$  satisfies the following semi-martingale representation theorem:

$$\mathcal{X}_t = \mathcal{X}_0 + \int_0^t \mathbf{Q} \mathcal{X}_u du + M_t$$

where  $M = \{M_t, t \in \mathcal{T}\}$  is a martingale with respect to the filtration generated by  $\mathcal{X}$  and  $\mathbf{Q}$  is the intensity matrix for  $N$  number of regimes,

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \ddots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{NN} \end{pmatrix}. \tag{2.1}$$

The risk-free rate  $\{r_t\}_{t \in \mathcal{T}}$ , risky asset return rate  $\{\alpha_t\}_{t \in \mathcal{T}}$  and volatility  $\{\sigma_t\}_{t \in \mathcal{T}}$  in the underlying asset dynamics are defined as:

$$\begin{aligned} r(t) &:= \mu(t, \mathcal{X}_t) = \langle r, \mathcal{X}_t \rangle = \sum_{i=1}^N r_i \langle \mathcal{X}_t, e_i \rangle, \\ \alpha(t) &:= \mu(t, \mathcal{X}_t) = \langle \alpha, \mathcal{X}_t \rangle = \sum_{i=1}^N \alpha_i \langle \mathcal{X}_t, e_i \rangle, \\ \sigma(t) &:= \sigma(t, \mathcal{X}_t) = \langle \sigma, \mathcal{X}_t \rangle = \sum_{i=1}^N \sigma_i \langle \mathcal{X}_t, e_i \rangle, \end{aligned}$$

where  $r := (r_1, r_2, \dots, r_N)$ ,  $\alpha := (\alpha_1, \alpha_2, \dots, \alpha_N)$  and  $\sigma := (\sigma_1, \sigma_2, \dots, \sigma_N)$  with  $\sigma_i > 0$  for all regimes  $i = 1, 2, \dots, N$ .

We use  $V_i$  to denote the objective function for regime  $i$ . Then the dynamic programming equation is

$$0 = (U_1(C, t) + \mu(t)U_2(L, t)) - \mu(t)V(W(t), t) - r(t)V(W(t), t) \tag{2.2}$$

$$+ V_W(W(t), t) [(\alpha - r)\pi(t)W(t) + r \cdot W(t) + Y - C(t) - P(t)] \tag{2.3}$$

$$+ \frac{1}{2} V_{WW}(W(t), t) \pi^2 W(t)^2 \sigma^2(t) + V_t(W(t), t) \tag{2.4}$$

$$+ \frac{\lambda(t)}{\mathcal{G}_2(t) - \mathcal{G}_1(t)} \int_{\mathcal{G}_1(t)}^{\mathcal{G}_2(t)} [V(W(t) + \pi(e^u - 1)W) - V(W)] du + \sum_j q_{ji} V_j(W(t), t), \tag{2.5}$$

where  $q_{ji}$  is the intensity rate from regime  $j$  to  $i$ .

We use the power utility function for consumption and bequests,

$$U_1(C(t)) = \frac{C(t)^\gamma}{\gamma}, \tag{2.6}$$

$$U_2(L(t)) = m(t)^{1-\gamma} \frac{L(t)^\gamma}{\gamma}, \tag{2.7}$$

where  $\gamma$  is the risk-aversion parameter and  $m(t)^{1-\gamma}$  is the discount function for bequests,

$$m(t) = e^{-\rho t/(1-\gamma)} \nu \int_t^\tau e^{-r(u-t)} du \tag{2.8}$$

and  $\nu$  is a constant that reflects the annuity level of an agent’s spouse or children compared to the current consumption amount.

From Richard (1975) and Purcal and Piggott (2008), the original objective function  $V$  has the assumed form,

$$V(W, t) = a(t) \frac{W^\gamma}{\gamma}. \tag{2.9}$$

Following Song et al. (2006), we can rewrite Eq. (2.9) as

$$V_i(W, t) = \sum_j \tilde{P}_{ji} a_j(t) \frac{W^\gamma}{\gamma}, \quad i = 1, 2, \dots, \tag{2.10}$$

where  $V_i(W, t)$  is the objective function for regime  $i$ ,  $\tilde{P}_{ji}$  is the transition probability of state  $j$  switching to state  $i$  and  $a_i(t)$  is the coefficient in the value function at time  $t$  for regime  $i$ .

Applying the first-order condition and substituting Eq. (2.9) into Eq. (2.5) in each regime, we have the optimal control variables

$$C^*(t) = \sum_i a_i(t)^{\frac{1}{\gamma-1}} W \cdot 1_{\{\mathcal{X}_t=i\}}, \tag{2.11}$$

$$\pi^*(t) = \frac{1}{(1-\gamma)\sigma^2(t)} \left[ \alpha(t) - r(t) + \frac{\lambda(t)}{\mathcal{G}_2(t) - \mathcal{G}_1(t)} \int_{\mathcal{G}_1(t)}^{\mathcal{G}_2(t)} G^{\gamma-1} (e^u - 1) du \right] \cdot 1_{\{\mathcal{X}_t=i\}}, \tag{2.12}$$

and

$$L^*(t) = \sum_i m(t) a_i(t)^{\frac{1}{\gamma-1}} W \cdot 1_{\{\mathcal{X}_t=i\}}, \tag{2.13}$$

where  $G = 1 + \pi^*(e^u - 1)$ . Equation (2.12) does not have the closed-form result. We need to use a numerical method to obtain the value of  $\pi$  for the given parameter values.

Given that the current regime is  $i$ , then substituting Eq. (2.9)–Eq. (2.13) into Eq. (2.5) and dividing it by  $W^\gamma/\gamma$ , we can have

$$\begin{aligned}
 0 = & \mu(t)m(t)a_i(t)^{\frac{\gamma}{\gamma-1}} + a_i(t)^{\frac{\gamma}{\gamma-1}} - (\mu(t) + \rho)a_i(t) + a_i'(t) \\
 & - \gamma(1 + \mu(t)m(t))a_i(t)^{\frac{\gamma}{\gamma-1}} + (r(t) + \mu(t))a_i(t)\gamma + (\alpha(t) - r(t))\pi^*a_i(t)\gamma \\
 & + \frac{1}{2}\sigma^2(t)(\pi^*)^2(\gamma - 1)a_i(t)\gamma + \frac{\lambda(t)a_i(t)}{\mathcal{G}_2(t) - \mathcal{G}_1(t)} \int_{\mathcal{G}_1(t)}^{\mathcal{G}_2(t)} (G^\gamma - 1)du + \sum_j q_{ji}a_j(t).
 \end{aligned}
 \tag{2.14}$$

By rearranging Eq. (2.14), we can have

$$\begin{aligned}
 & a_i(t)^{\frac{\gamma}{\gamma-1}}[(\gamma - 1)(1 + \mu(t)m(t))] \\
 & = a_i'(t) + a_i(t) \left[ -\mu(t) - \rho + (r + \mu(t))\gamma + (\alpha(t) - r(t))\gamma\pi^* \right. \\
 & \left. + \frac{1}{2}\sigma^2(t)(\pi^*)^2(\gamma - 1)\gamma + \frac{\lambda(t)}{\mathcal{G}_2(t) - \mathcal{G}_1(t)} \int_{\mathcal{G}_1(t)}^{\mathcal{G}_2(t)} (G^\gamma - 1)du + \sum_j q_{ji} \right].
 \end{aligned}
 \tag{2.15}$$

As we have the dynamic programming equation as Eq. (2.15) for each regime, we end up with a system of equations. Through the numerical schemes from Song et al. (2006), we can solve Eq. (2.15) and obtain the numerical solution for coefficient  $a_i(t)$ . The consumption, investment and insurance premiums can then be calculated.

### 3 Numerical Results

#### 3.1 Parameter Values

In this paper, the American survival probability and force of mortality are obtained from the American 2010 life table for males (Human Mortality Database N.d.). We calibrate our parameters to the American data to obtain numerical results from the starting age  $t = 65$  to the maximum age  $\omega = 109$  (Table 1).

**Table 1** Parameters used in the numerical results

$t=65$	$\tau=109$	$\rho r_1 = 0.04187$
$W = \text{US}\$522,500$	$\gamma - 2$ or $-0.8$	$\nu$ 2/3 or 1
$\mathcal{G}_1 = -0.5$	$\mathcal{G}_2 = 0.5$	$\lambda$ 0, 0.2, 0.4 and 0.6
$r_1 = 0.04187$	$\alpha_1 = 0.0592$	$\sigma_1 = 0.1853$
$r_2 = 0.03$	$\alpha_2 = 0.04$	$\sigma_2 = 0.3$
$r_3 = 0.025$	$\alpha_3 = 0.03$	$\sigma_3 = 0.4$

We calculate the regime 1 risk-free rate,  $r_1$ , by using the 1-year yield rate (from 1990 to 2010) from the US Department of the Treasury (2015). The agent’s time preference is assumed to be the same as the risk-free rate,  $\rho = r_1$  for simplicity. Meanwhile, the regime 1 risky asset return rate,  $\alpha_1$ , and volatility,  $\sigma_1$  are calibrated to Standard & Poor’s (S&P) 500 data from 1990 to 2010.

We also adopt the average income,  $Y = \$52250$  from Noss (2013). Agents in our model are assumed to have a total wealth of  $10Y$  from previous savings and have no future income. The risk-aversion parameter  $\gamma$  is set to be  $-2$  for the jump case and  $-0.8$  for the regime switching case. Here, the  $\nu$  value illustrates the annuity level left for the surviving spouse or children which is an indicator for a bequest motive. From Purcal and Piggott (2008), we set  $\nu$  to be two-thirds for a low bequest motive, that is, agents will only leave two-thirds of their current consumption level for the surviving spouse or children. As we want to further study the effects from a higher bequest motive, we also calculate our result for  $\nu = 1$  for a high bequest motive.

For the jump case, we assume the values of  $\mathcal{G}_1$  and  $\mathcal{G}_2$  to be  $-0.5$  and  $0.5$ , respectively. To test the jump effects, different values are tried for the jump frequency, that is,  $\lambda = 0, 0.2, 0.4$  and  $0.6$ . When  $\lambda = 0$ , this means that there is no jump in the risky asset.

To demonstrate the numerical results for the case with jumps and regime switching, we assume there are three regimes: regime 1 “good”, regime 2 “okay” and regime 3 “bad”. We assume that agents initially start in the good regime. For regime 2 and 3, we assume the following risk-free rates, risky asset return rates and volatility rate, regime 2:  $r_2 = 0.03$   $\alpha_2 = 0.04$   $\sigma_2 = 0.3$ ; regime 3:  $r_3 = 0.025$   $\alpha_2 = 0.03$   $\sigma_2 = 0.4$ . With these preset values, a worse economic environment is described by the lower risk-free rate, lower risky asset return rate and higher volatility. We assume the preset intensity rate for switching is as follows:

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}. \tag{3.16}$$

### 3.2 Case with Jumps

Here, we compare the cases with different levels of jump frequency and bequest motive, that is,  $\lambda = 0, 0.2, 0.4$  or  $0.6$  and  $\nu = 2/3$  or  $1$ . Specifically, we calculate the expected proportions of wealth, consumption, insurance premium and risky asset investment for each  $\lambda$  and  $\nu$  value, as shown in Figs. 1, 2, 3 and 4.

As our model is to study the post-retirement period, the wealth level will decline when agents are ageing. From Figs. 1 and 2, we observe that a higher jump frequency can result in less wealth for both levels of bequest motive. This phenomenon can be linked with variations in the proportion of wealth invested in risky assets, that is,  $\pi$  due to different jump frequencies. Table 2 shows that agents would reduce their

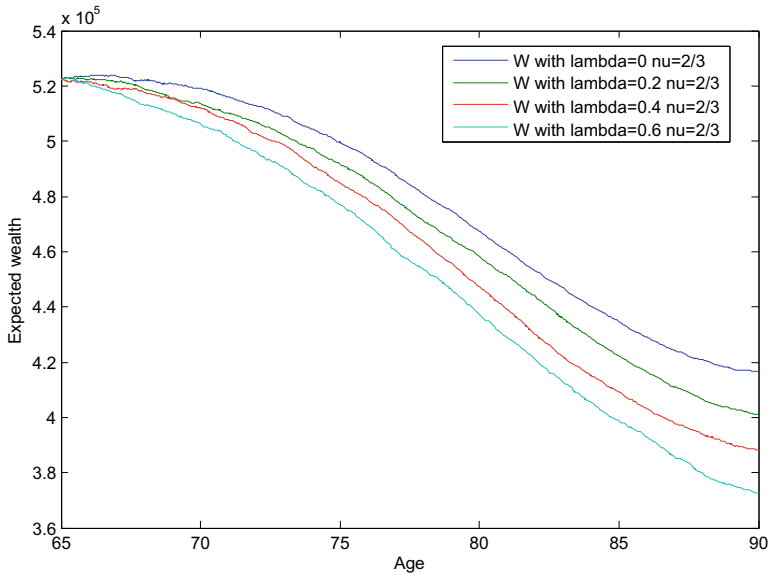


Fig. 1 Expected wealth with jumps  $\gamma = -2, \nu = 2/3, \lambda = 0, 0.2, 0.4$  or  $0.6$

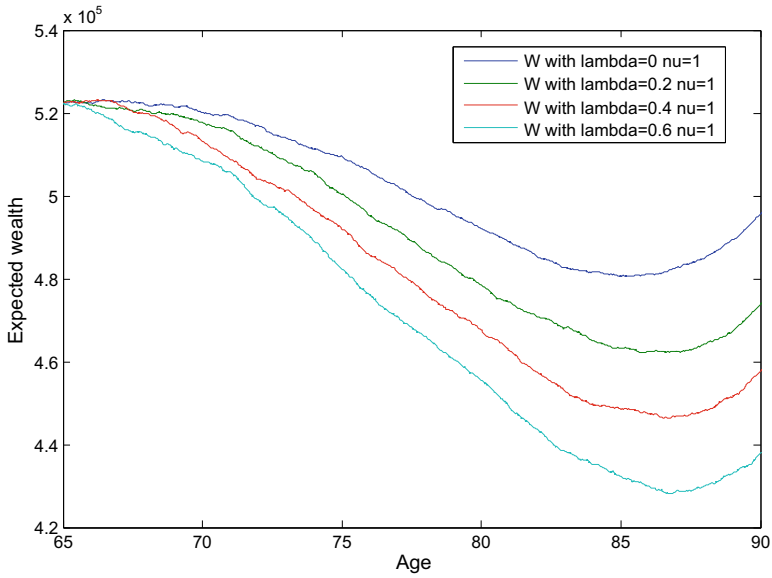


Fig. 2 Expected wealth with jumps  $\gamma = -2, \nu = 1, \lambda = 0, 0.2, 0.4$  or  $0.6$



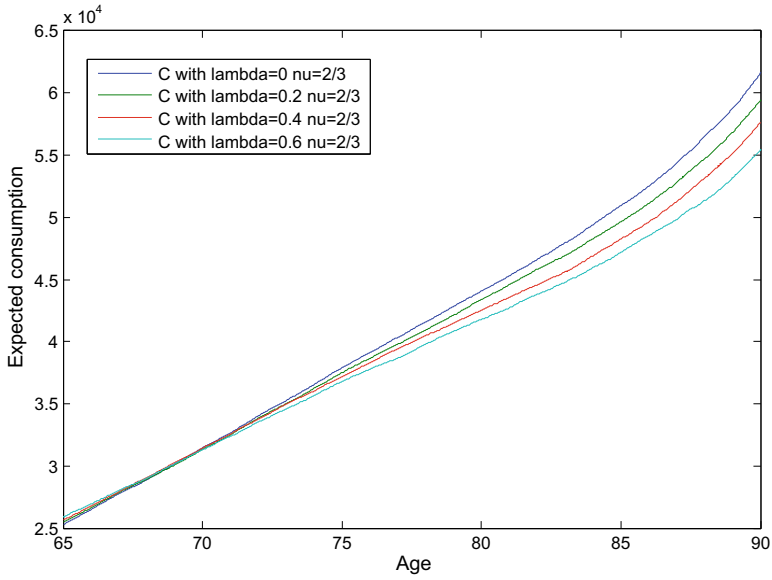


Fig. 3 Expected consumption with jumps  $\gamma = -2$ ,  $\nu = 2/3$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$

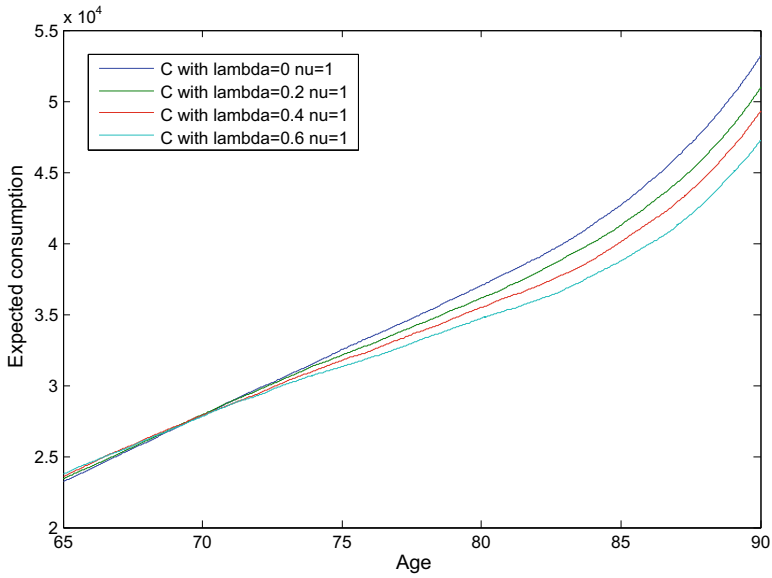


Fig. 4 Expected consumption with jumps  $\gamma = -2$ ,  $\nu = 1$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$

**Table 2** Investment proportion in risky assets for different jump frequencies

$\lambda$	$\pi$ with $\nu = 2/3$ or 1
0	0.1680
0.2	0.1668
0.4	0.1662
0.6	0.1659

exposure to risky assets when jumps are more frequent. Comparing the results in Figs. 1 and 2, agents with a higher bequest motive are willing to hold more wealth.

From Figs. 3 and 4, the expected increase in consumption level in line with increasing age for different  $\lambda$  values can be found for both levels of bequest motive. With a higher jump frequency, that is, a higher  $\lambda$  value, the expected consumption level is lower in line with increasing age. We also notice that agents have a propensity to consume less when they have a higher bequest motive.

In our model, the insurance premium is the indicator for life insurance or annuity demand. A positive insurance premium, that is  $P$ , indicates agent's demand for life insurance, while a negative insurance premium indicates the demand for annuitisation. In Fig. 5, we have all negative insurance premiums which illustrate agents' annuitisation intention. With more frequent potential jumps in the risky assets, agents reduce their annuitisation amount due to their lower level of wealth. However, as shown in Fig. 6, with a higher bequest motive, agents' demand for life insurance starts a few years after retirement for all market environments, with the higher level of demand corresponding to the higher  $\lambda$  value.

### 3.3 Case with Jumps and Regime Switching

We conduct another numerical calculation for the case with jumps and regime switching. To study the effects of regime switching and a bequest motive, we calculate the numerical result with or without regime switching for different levels of bequest motive, as shown in Figs. 7, 8, 9 and 10, when there are jumps in the financial market.

From Figs. 7 and 8, agents with both low and high bequest motives are found to have less consumption motivation when there are additional risks present due to regime switching. Bequest motive effects are detected here, as agents tend to consume less for higher  $\nu$  value.

Similarly, as shown in Figs. 9 and 10, agents express lower annuitisation intension when they are faced with additional risks from regime switching. Furthermore, agents would reduce their annuitisation intension if they have a higher level of bequest motive.

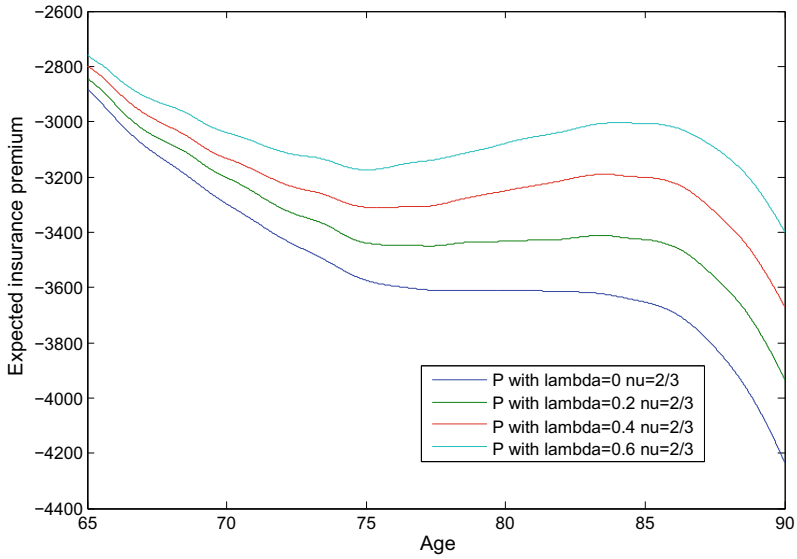


Fig. 5 Expected insurance premium with jumps  $\gamma = -2, \nu = 2/3, \lambda = 0, 0.2, 0.4$  or  $0.6$

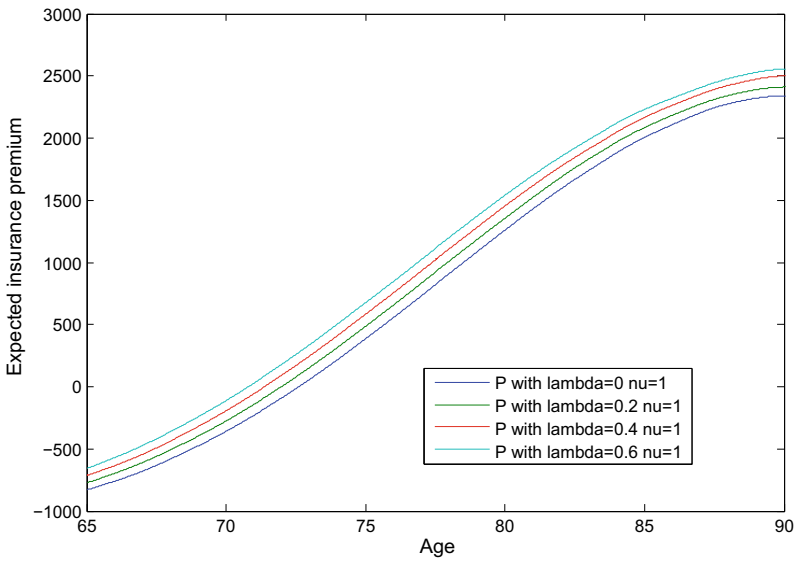
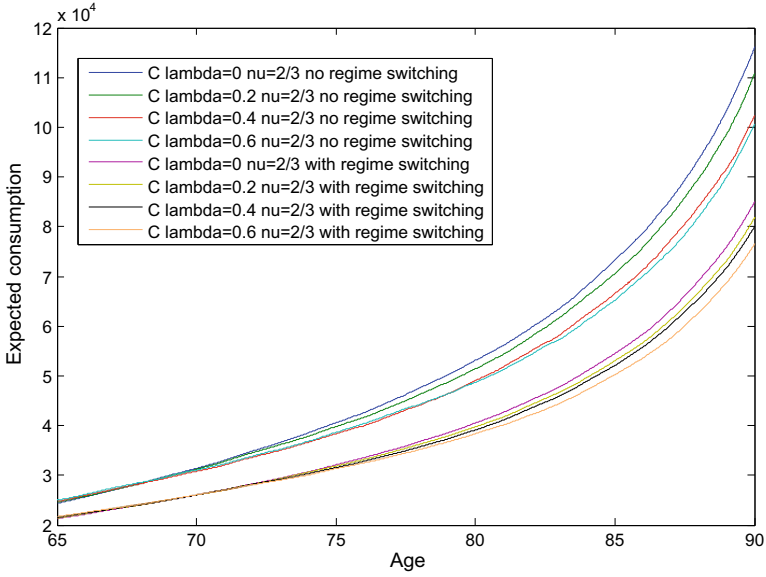
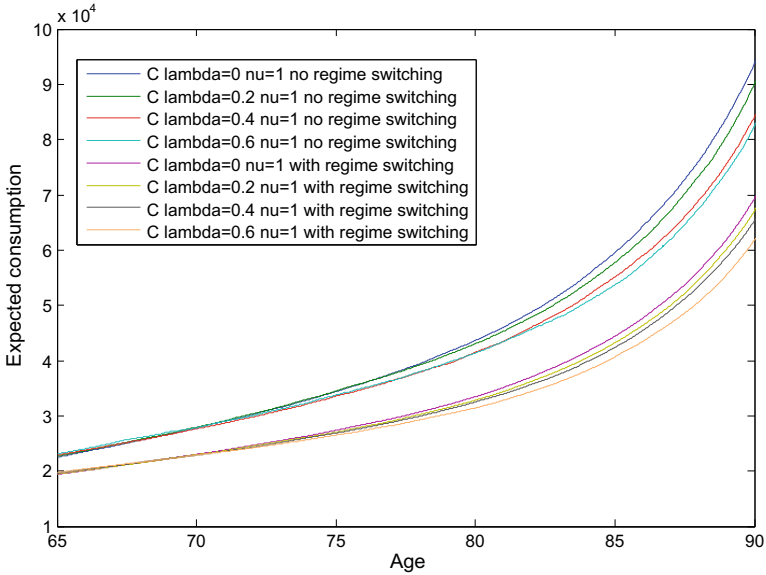


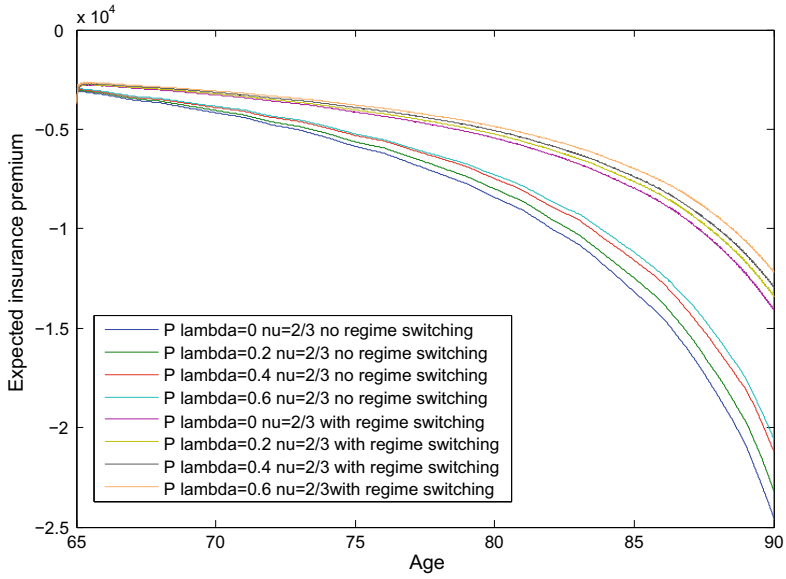
Fig. 6 Expected insurance premium with jumps  $\gamma = -2, \nu = 1, \lambda = 0, 0.2, 0.4$  or  $0.6$



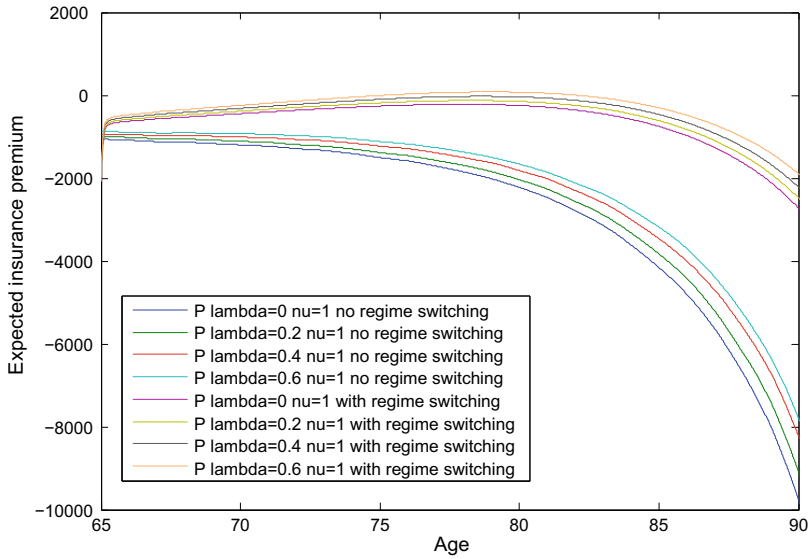
**Fig. 7** Expected consumption with jumps and regime switching  $\gamma = -0.8$ ,  $\nu = 2/3$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$



**Fig. 8** Expected consumption with jumps and regime switching  $\gamma = -0.8$ ,  $\nu = 1$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$



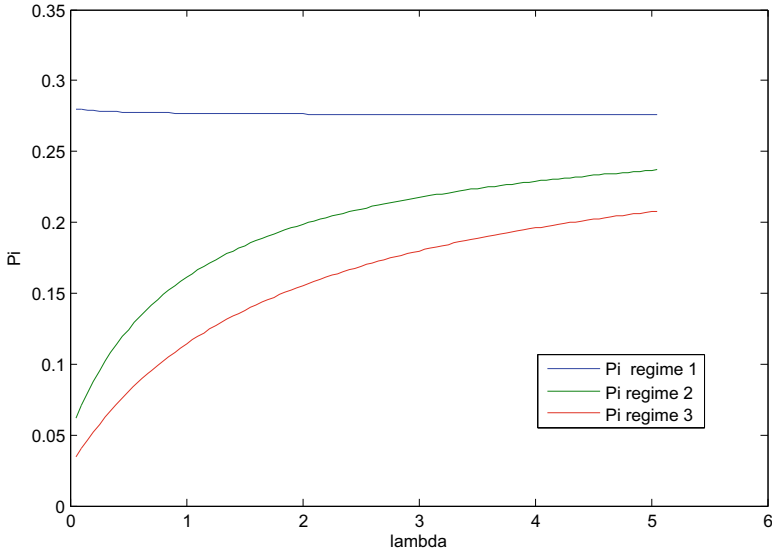
**Fig. 9** Expected insurance premium with jumps and regime switching  $\gamma = -0.8$ ,  $\nu = 2/3$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$



**Fig. 10** Expected insurance premium with jumps and regime switching  $\gamma = -0.8$ ,  $\nu = 1$ ,  $\lambda = 0, 0.2, 0.4$  or  $0.6$

**Table 3** Investment proportion in risky assets with jumps and regime switching

$\lambda$	$\pi$ with jumps	$\pi$ with jumps and regime switching
0	0.2800	0.1256
0.2	0.2786	0.1438
0.4	0.2779	0.1579
0.6	0.2774	0.1691



**Fig. 11**  $\pi$  value for three regimes  $\gamma = -0.8, \lambda = 0, 0.05, 0.1, \dots, 4.95, 5$

We also calculate the proportion of agents’ wealth invested in risky assets in the economic environment with jumps and in the economic environment with jumps and regime switching in Table 3. From the  $\pi$  value, compared to the case that only has jumps, agents tend to reduce their exposure to risky assets when regime switching occurs. With the increasing probability of the potentially worse economic environment that is captured by more frequent jumps, agents would like to increase their investment proportion. This result is coincident with the prospect theory (Kahneman and Tversky 1979), in which agents are believed to take more risks involving a higher probability of losses.

As shown in Fig. 11, we identify some interesting aspects from the  $\pi$  value. With different  $\lambda$  values, we then calculate the  $\pi$  value for each regime by numerically solving Eq. (2.12) in Fig. 11. Based on Eq. (2.12) and Fig. 11,  $\pi$  is dependent on the  $\sigma$  value which can either increase or decrease with an increase in the  $\lambda$  value.

## 4 Conclusion

In this paper, we extend Richard's model (Richard 1975) to examine agents' investment behaviour during changes in the economic state. Using our model, we study agents' post-retirement investment behaviour in relation to risky assets when economic dynamics are described by jumps and regime switching.

Based on our numerical results for the case with jumps, agents' behaviour deviates when they experience changes in the economic state. When agents detect potential future jumps, they will reduce their exposure to risky assets which will result in lower wealth, consumption and annuitisation. This type of behaviour will become more substantial with a higher jump frequency. When the agents' bequest motive is higher, they might further reduce their consumption and annuitisation. In fact, agents will seek life insurance, if their wealth cannot cover the legacy amount.

When there are both potential jumps and regime switching in the market, agents have the tendency to further reduce their consumption and annuitisation compared to when the market has jumps only. With a higher bequest motive, this trend is more obvious. However, according to our calculations, agents would plan to enhance their exposure to risky assets when jumps are more frequent.

We can conclude that agents' behaviour will be different when they are in the presence of risks from a changing economic state. With reduced wealth, they will be more conservative which is shown in their reduced consumption and annuitisation. From agents' behaviour, we find that having a bequest motive has negative effects on consumption and annuitisation. However, in our model for different volatility values, agents can either exhibit their preference for risk taking or risk aversion when the market is more volatile. Specifically, agents who are in regime 1, that is in a good economic state, reduce their proportion of wealth invested in risky assets along with increasing jump frequency. Meanwhile, agents who are in regime 2 and 3, that is in a worse economic state, raise their proportion of wealth invested in risky assets along with increasing jump frequency. When agents are confronting a market that has a high degree of variation, it is more common for us to anticipate a risk-aversion behaviour rather than the risk-taking behaviour. However, in our model, jump is assumed to be negative and positive. Agents in regime 2 and 3 are confronting lower return rate but higher volatility. Those agents would then adopt risk-taking behaviours due to the tendency for positive jump and probability of switching to a better regime. On the other hand, agents in regime 1, who could have fears for potential downgrade movement, would like to be more conservative and risk-averse.

Other explanations for the risk-taking tendency taken by agents can be found in various existing literatures, such as Kahneman and Tversky (1979), and Shum and Faig (2006). The prospect theory model that is studied in Kahneman and Tversky (1979) indicates agents can overweight the small probability of an outcome. Hence, agents might increase the investment in risky assets when the market is in a worse condition (lower return rate but higher volatility), as they overweight the probability of a high return rate. The increasing proportion of risky assets can also be explained by the retirement saving target (Shum and Faig 2006). The empirical study in Shum

and Faig (2006) states retirement saving motive has a positive impact on agents' shares investment. As agents in our model are assumed to have the bequest motive, they also have the retirement saving target. Therefore, to pursuit the saving target, agents would like to hold more risky assets in a volatile market.

## References

- Bateman H, Islam T, Louviere J, Satchell S, Thorp S (2011) Retirement investor risk tolerance in tranquil and crisis periods: experimental survey evidence. *J Behav Finance* 12(4):201–218
- Bollen NPB (1998) Valuing options in regime-switching models. *J Deriv* 6(1):38–49
- Brown JR (2001) Private pensions, mortality risk, and the decision to annuitize. *J Public Econ* 82(1):29–62
- Cont R, Tankov P (2004) Financial modelling with jump processes, vol 133. Chapman & Hall/CRC
- Elliott RJ, Aggoun L, Moore JB (1994) Hidden Markov models: estimation and control. Springer
- Elliott RJ, Siu TK, Chan L, Lau JW (2007) Pricing options under a generalized Markov-modulated jump-diffusion model. *Stoch Anal Appl* 25(4):821–843
- Gupta A, Murray W (2003) How to spend and invest retirement savings. *Ann Oper Res* 124(1–4):205–224
- Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econom: J Econom Soc* 57(2):357–384
- Hanson FB (2007) Applied stochastic processes and control for jump-diffusions: modeling, analysis, and computation, vol 13. SIAM
- Heaton J, Lucas D (2000) Portfolio choice in the presence of background risk. *Econ J* 110(460):1–26
- Human Mortality Database (N.d.) U.S.A males life table. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). [www.mortality.org](http://www.mortality.org) or [www.humanmortality.de](http://www.humanmortality.de). Accessed 15 Jan 2016 on <http://www.mortality.org>
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econom: J Econom Soc* 47(2):263–291
- Merton RC (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. *Rev Econ Stat* 51(3):247–257
- Merton RC (1976) Option pricing when underlying stock returns are discontinuous. *J Financ Econ* 3(1):125–144
- Noss A (2013) Household income: 2012. United States Census Bureau, US Department of Commerce 12(2)
- Purcal S, Piggott J (2008) Explaining low annuity demand: an optimal portfolio application to Japan. *J Risk Insur* 75(2):493–516
- Purcal S, Wang TH (2005) Optimal consumer behaviour in a jump-diffusion environment. University of New South Wales
- Richard SF (1975) Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model. *J Financ Econ* 2(2):187–203
- Shum P, Faig M (2006) What explains household stock holdings? *J Bank Finance* 30(9):2579–2597
- Song QS, Yin G, Zhang Z (2006) Numerical methods for controlled regime-switching diffusions and regime-switching jump diffusions. *Autom: J IFAC* 42(7):1147–1157
- Sotomayor LR, Cadenillas A (2009) Explicit solutions of consumption-investment problems in financial markets with regime switching. *Math Finance* 19(2):251–279
- US Department of the Treasury (2015) Daily treasury yield curve rates. Washington, United States
- Wu L, Kwok Y-K (1997) A front-fixing finite difference method for the valuation of American options. *J Financ Eng* 6(4):83–97
- Zhang Q, Guo X (2004) Closed-form solutions for perpetual American put options with regime switching. *SIAM J Appl Math* 64(6):2034–2049



# Prediction of Claim Probability with Excess Zeros



Aslıhan Şentürk Acar

**Abstract** Non-life insurance pricing is based on two components: claim severity and claim frequency. These components are used to estimate expected pure premium for the next policy period. Generalized linear models (GLM) are widely preferred for the estimation of claim frequency and claim severity due to the ease of interpretation and implementation. Since GLMs have some restrictions such as exponential family distribution assumption, more flexible Machine Learning (ML) methods are applied to insurance data in recent years. ML methods use learning algorithms to establish relationship between the response and the predictor variables as an intersection of computer science and statistics. Because of some insurance policy modifications such as deductible and no claim discount system, excess zeros are usually observed in claim frequency data. In the presence of excess zeros, prediction of claim probability can be a good alternative to the prediction of claim numbers since positive numbers are rarely observed in the portfolio. Excess zeros create imbalance problem in the data. When the data is highly imbalanced, predictions will be biased toward majority class due to the priors and predicted probabilities may be uncalibrated. In this study, we are interested in claim occurrence probability in the presence of excess zeros. A Turkish motor insurance dataset that is highly imbalanced is used for the case study. Ensemble methods that are popular ML approaches are used for the probability prediction as an alternative to logistic regression. Calibration methods are applied to predicted probabilities and results are compared.

**Keywords** Claim probability · Imbalanced data · Non-life insurance · Machine learning

---

A. Ş. Acar (✉)

Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey

e-mail: [aslihans@hacettepe.edu.tr](mailto:aslihans@hacettepe.edu.tr)

## 1 Introduction

Insurance companies guarantee to compensate policyholder against unpredictable losses during a certain time period by charging premium for the assurance. Basic objective of ratemaking is to determine fair premiums for the policyholders that have different characteristics. For this purpose, actuaries use statistical models and rating factors to determine premiums. Statistical approach depends on the observed data in a certain accounting year. Observed responses of each policy can be only aggregate losses, both total claim numbers and aggregate losses or detailed information for each claim event (Frees et al. 2014). According to actuarial equivalence principle, pure premium is equal to expected total claim size that depends on two components: expected claim frequency and expected claim severity. When observed responses are only aggregate losses in a policy period, we need two components to estimate pure premium: estimates of total claim size and claim probability.

Non-life insurance data have some characteristics such as excess of zeros in claim numbers due to NCD system and deductible modification. Excess of zeros leads to imbalanced data structure. In such a case, predictive models perform poorly because of few instances of minority class and they classify most of the observations as majority class. When the data is highly imbalanced, predictions will be biased toward majority class due to the priors (Kuhn and Johnson 2013). Since we are interested in claim (occurrence) probability that constitutes minority class, we have to deal with imbalance structure of data and the bias to get accurate predictions. To deal with imbalanced data, a common approach is to apply resampling methods (He and Garcia 2009; Japkowicz and Stephen 2002). Various resampling methods are used to rebalance the data such as oversampling, undersampling, and synthetic minority oversampling technique. In addition to resampling methods, feature selection, cost sensitive learning, and hybrid ensemble learning methods are other alternatives to deal with class imbalance (Guo et al. 2008).

From the insurance perspective, zero-inflated and hurdle models are used to deal with excess of zeros (Yip and Yau 2005; Boucher et al. 2007). These models include a component for structural zeros that has to be estimated using generally logistic regression (LR). Although LR is easy to understand and implement, it is constrained to a specified form that creates consistency problems when model is not correctly specified. As an alternative to LR, flexible nonparametric machine learning algorithms are used in many studies in recent years. Although LR can directly predict calibrated probabilities due to the optimization of log loss, some ML methods don't produce calibrated probabilities specially for imbalanced data and they need calibration (Fernández et al. 2018). Uncalibrated probabilities may induce bias in probability scores. If probabilities are calibrated they will represent the likelihood of true classes. Platt scaling (Platt 1999) and isotonic regression (Zadrozny and Elkan 2001) are two popular calibration methods in the literature (Niculescu-Mizil et al. 2012).

In this study, we are interested in claim occurrence probability in the presence of excess zeros. To the best of our knowledge there are very few studies in actuarial literature that use ML methods for the prediction of claim probability. Dal Pozzolo

(Pozzolo 2010) used decision trees, random forest, Naïve Bayes, K-nearest neighbors, neural networks (NN), support vector machine (SVM), and linear discriminant analysis to classify claims whether they are greater than zero or not using claim probability estimates based on different thresholds. Frempong et al. (Frempong et al. 2017) used decision trees to predict probability of making a claim. Pijl (Tim Pijl 2017) used decision trees, random forest (RF), LR, and SVM to predict probability of issuing a claim.

We aim to compare predictive performances of LR and ensemble methods to predict claim probability in the presence of excess zeros. Calibration methods are applied to predicted probabilities and the results are compared using Brier score (BS) (Glenn 1950).

## 2 Methods

### 2.1 Logistic Regression

Response variable is binary ( $Y = 0$  or  $Y = 1$ ) in classification tasks. Assuming  $p = P(Y = 1)$ , logistic regression equation is expressed as

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = x'\beta \quad (1)$$

where  $x$  is the vector of predictors and  $\beta$  is the vector of regression parameters. Predictions of LR are probabilities of binary event.

### 2.2 Bagging

Ensemble methods are designed to improve the predictive performance of decision trees. These methods compile the information related to the predictions of base models. They have no distributional restrictions and they handle interactions between variables easily. Bagging (bootstrap aggregating) is the simplest ensemble method in that bootstrap samples are drawn randomly from the study sample with replacement to reduce the variance. Different training samples are created using bootstrap, generally decision trees are chosen as base learning algorithm and the predictions are averaged over bootstrap samples (Breiman 1996). Predicted probability of binary event for a unit is obtained as the ratio of units that have the event among all units in related subset (Austin et al. 2013).

### 2.3 *Random Forest*

Random forest approach (Breiman 2001) also uses bootstrap samples similar as bagging but considers binary splits of tree on a random sample of predictor variables instead of all candidate predictor variables to decorrelate the trees and increase the accuracy (Austin et al. 2013; James et al. 2013). When the number of randomly selected predictors is equal to the total number of predictors, random forest algorithm reduces to the bagging algorithm.

### 2.4 *Boosting*

Boosting works in similar way with bagging but with boosting method each tree uses the information from previous tree and trees are grown sequentially by applying weak learner to the reweighted data (James et al. 2013). Objective is to reduce the error of a weak learner and get a strong predictor (Freund and Schapire 1996). There are several boosting algorithms in the literature but generalized boosted model (GBM) (Friedman 2001) is used in this study.

## 3 *Case Study*

### 3.1 *About Dataset*

Case study is implemented using motor insurance dataset from an insurance company in Turkey. There are 376,719 individual automobile policies in the portfolio. All policies are started or renewed in 2010 and each policy has 1 year of exposure. Very few policies had more than one claim during the policy period. Frequency of claim numbers is given in Table 1.

There are only four individuals that have reported four claims and %0.15 of policies had more than one claim during 1-year policy period. Data is highly imbalanced since %95.43 of policyholders did not report any claim during the policy year. Therefore, we prefer to model claim probability instead of claim numbers. Predictor

**Table 1** Frequency of claim numbers

Number of claims	Number of records
0	359,487
1	16,660
2	540
3	28
4	4

variables are age of policyholder (18–90), gender of policyholder (0:female, 1:male), province in which number plate of vehicle is registered, age of vehicle (0–64), and horse power of vehicle. Based on frequency information, we cut age of policyholder at 90. Provinces are clustered into seven clusters using 2010 year accident statistics that are published by Turkish Statistical Institute. Statistics related to the continuous predictors are given in Table 2. Frequencies related to categorical variables are given in Table 3.

As can be seen from Table 3, most of the policyholders are the males.

**Table 2** Statistics of continuous predictor variables

Claim (1) No Claim (0)	Min	1st Qu	Median	Mean	3rd Qu	Max
Age of policyholder						
0	18	32	40	42.13	50	90
1	18	31	39	40.74	49	90
Age of vehicle						
0	0	7	13	13.23	18	88
1	0	6	12	12.57	17	62
Horse power						
0	20	75	80	88.69	100	1001
1	26	75	80	89.86	100	445

**Table 3** Frequencies of categorical variables (percentage)

Province	No claim	Claim
0	42,845 (%11)	3030 (%0.8)
1	71,641 (%19)	2897 (%0.8)
2	26,795 (%7)	965 (%0.3)
3	76,647 (%20)	3628 (%1)
4	53,139 (%14)	2550 (%0.7)
5	31,622 (%8)	1860 (%0.5)
6	56,798 (%15)	2302 (%0.6)
Gender	No claim	Claim
Female	48,747 (%13)	2761 (%1)
Male	310,740 (%82)	14,471 (%4)

**Table 4** LR estimation results

Parameter	Estimate	Std. Error	z value	Pr(> z )
intercept	-2.153	0.054	-40.196	< 2e-16
Male	-0.141	0.024	-5.868	4.41e-09
Age	-0.008	0.003	-17.204	< 2e-16
province1	-0.524	0.030	-17.012	< 2e-16
province2	-0.698	0.043	-16.252	< 2e-16
province3	-0.357	0.029	-12.474	< 2e-16
province4	-0.364	0.031	-11.678	< 2e-16
province5	-0.181	0.034	-5.276	1.32e-07
province6	-0.520	0.032	-16.254	< 2e-16
vage	-0.007	0.001	-5.888	3.92e-09
horse power	0.0004	0.0003	1.334	0.182

### 3.2 Analysis

Data is randomly partitioned into two parts: training dataset (%80) and test dataset (%20). Models are fitted to training data and validated using test data. No interaction and nonlinear terms are used in LR. Parameter estimates of LR are given in Table 4. From Table 4, we can say that all predictor variables except horse power are statistically significant at %95 confidence level.

For the ease of computation, fivefold cross-validation (CV) method is used for tuning the hyperparameters of ML methods. There is no hyperparameter for bagging algorithm. A tree is constructed for each of drawn 50 bootstrap samples. Random forest algorithm has three hyperparameters: the number of predictors selected at each split, split rule, and minimal node size (Wright et al. xxxx). Finally, generalized boosted regression model has four hyperparameters: number of trees, maximum depth of trees, learning rate, and minimum observation number in terminal nodes.

Statistics related to the predicted probabilities of policyholders in test dataset are given in Table 5. Interval of probability predictions of LR and GBM is too narrow that reflect the imbalance structure of the dataset. We can easily say that predictions

**Table 5** Statistics related to the predicted claim probabilities

Statistics	LR	Bagging	RF	GBM
Min	0.0172	0.0000	0.0015	0.0304
1st Qu	0.0380	0.0000	0.0269	0.0414
Median	0.0424	0.0000	0.0383	0.0448
Mean	0.0456	0.0326	0.0466	0.0456
3rd Qu	0.0510	0.0000	0.0573	0.0490
Max	0.0980	0.9500	0.4471	0.0698

**Table 6** Brier score values of predictive models

Model	BS
LR	0.0439
Bagging	0.0523
RF	0.0445
GBM	0.0440

are biased toward majority class. Bagging predicts too many zeros (median is zero) compared to other methods because it is a weaker learner compared to RF and GBM.

To deal with imbalanced structure of dataset, we planned to use resampling methods to investigate effect of resampling on predictive performance. Racing algorithm (Birattari et al. 2002) of unbalanced R package (Pozzolo and Caelen xxxx) is applied to select the best resampling method. According to the result, none of the resampling algorithm is suitable for our dataset. Eventually, we did not apply any resampling scheme.

We compared predictive performance of candidate models using BS that is the mean squared loss between the predicted probabilities and actual responses. Lower BS means better predictive performance. BS values of each prediction method related to test dataset are given in Table 6.

According to Table 6, LR and GBM have best predictive performance with lowest BS values. But there is a very little difference with RF method. Bagging performs worse as expected from prediction values given in Table 5.

We used Platt scaling and isotonic regression for the calibration of probabilities predicted. In Platt scaling, a sigmoid function is used for the probabilities and gradient descent algorithm is used to find two parameters of sigmoid function (Platt 1999). Isotonic regression approach fits a nonparametric monotonic regressor. It minimizes the squared error between the true class labels and the outputs (Zadrozny and Elkan 2002). BS values after calibration methods are given in Table 7.

As seen from Table 7, both calibration methods are not effective on the predictive performances. Specially, with isotonic regression, BS values increased in bagging and GBM methods. We can conclude that calibration methods did not work for this highly imbalanced dataset.

**Table 7** BS values after calibration methods

Calibration methods	LR	RF	Bagging	GBM
Platt scaling	0.04391	0.04399	0.04400	0.04393
Iso Reg	0.04422	0.04424	0.25620	0.16706

## 4 Conclusion

Claim probability is an important measure about the uncertainty of claim occurrence. It also constitutes one part of two-part models such as ZIP that is frequently used for claim frequency modeling. In this study, main objective is to compare predictive performance of logistic regression and ensemble methods for the prediction of claim probability in presence of excess zeros. A Turkish motor insurance dataset is used for the case study. According to case study results, predicted probabilities were biased to majority class (zero) and calibration methods did not improve predictive performance of the methods based on Brier score values. Another result is that RF and GBM performed similar predictive performance with logistic regression. Bagging method performed worst among all predictive models since it neither has random variable selection nor works sequentially like RF and GBM to increase accuracy. Consequently, ML methods are good alternatives to classical approaches for the prediction.

As a future study, more complex ML methods such as neural networks can be used for the prediction of claim probability. Another idea is the comparison of claim severity\*frequency approach with the claim probability\*total claim size approach for the prediction of total claim amount.

## References

- Frees EW, Derrig RA, Meyers G (2014) Predictive modeling applications in actuarial science. Cambridge University Press, p 565
- Kuhn M, Johnson K (2013) Applied predictive modelling, vol 26. Springer
- He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 21(9):1263–1284
- Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. *Intell Data Anal* 6(5):429–449
- Guo X, Yin Y, Dong C, Zhou G (2008) On the class imbalance problem. *IEEE Conf Publ* 4:192–201
- Yip KCH, Yau KKW (2005) On modeling claim frequency data in general insurance with extra zeros. *Insur Math Econ* 36(2):153–163
- Boucher JP, Denuit M, Guillén M (2007) Risk classification for claim counts: a comparative analysis of various zeroinflated mixed poisson and hurdle models. *North Am Actuar J* 11(4):110–131
- Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets, vol 11. Springer, Berlin
- Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 10(3):61–74
- Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and Naive Bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine learning [Internet]. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 609–616. (ICML '01). Available from: <http://dl.acm.org/citation.cfm?id=645530.655658>
- Niculescu-Mizil A, Caruana RA (2012) Obtaining calibrated probabilities from boosting. Jul 4 [cited 2021 May 29]; Available from: <https://arxiv.org/abs/1207.1403v1>
- Pozzolo AD (2010) Comparison of data mining techniques for insurance claim prediction [Master of Science]. University of Bologna



- Frempong NK, Nicholas N, Boateng MA (2017) Decision tree as a predictive modeling tool for auto insurance claims. *Int J Stat Appl* 7(2):117–120
- Tim P (2017) A framework to forecast insurance claims [Master of Econometrics and Management Science]. Erasmus University Rotterdam
- Glenn W (1950) Brier, verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3
- Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
- Austin PC, Tu JV, Ho JE, Levy D, Lee DS (2013) Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol* 66(4):398–407
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning [Internet], vol 6. Springer. Available from: <https://doi.org/10.1007/978-1-4614-7138-7.pdf>
- Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In 1996. pp 148–56
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232
- Marvin NW, Wager S, Probst P (2018) “ranger” package
- Birattari M, Stützle T, Paquete L, Varrentrapp K (2002) A racing algorithm for configuring meta-heuristics. In: Proceedings of the 4th Annual Conference on Genetic and evolutionary computation [Internet]. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 11–18. (GECCO’02)
- Pozzolo AD, Caelen O, Bontempi G (2015) Package “unbalanced.”
- Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In 2002 [cited 2021 Jun 4]. Available from: <https://doi.org/10.1145/775047.775151>

# Risk Classification in Nonlife Insurance Premium Ratemaking



Amela Omerašević and Jasmina Selimović

**Abstract** The aim of this research is to explore and analyze the benefits of risk classification methods by using data mining techniques on premium ratemaking in nonlife insurance. We rely on generalized linear models (GLMs) framework for nonlife premiums and examine the impact of specific data mining techniques on classifications of risk in motor hull insurance in Bosnia and Herzegovina. We study this relationship in an integrated framework considering a standard risk model based on the application of Poisson GLM for claims frequency estimate. Although GLM is a widely used method to determine insurance premiums, improvements of GLM by using the data mining methods identified in this paper may solve practical challenges for the risk models. The application of the data mining method in this paper aims to improve the results in the process of nonlife insurance premium ratemaking. The improvement is reflected in the choice of predictors or risk factors that have an impact on insurance premium rates. The following data mining methods for the selection of prediction variables were investigated: forward stepwise and neural networks. We provide strong and robust evidence that the use of data mining techniques influences premium ratemaking in nonlife insurance.

**Keywords** Ratemaking · Forward stepwise · Neural networks · Nonlife insurance · GLM

## 1 Introduction

In the competitive national economies, in developed countries, the insurance premium that the insured pays in exchange for the risk transfer to the insurer is an extremely important aspect of insurance. Calculation of the premium, as the price

---

A. Omerašević  
Uniqa osiguranje d.d, Sarajevo, Bosnia and Herzegovina

J. Selimović (✉)  
School of Economics and Business, University of Sarajevo, Sarajevo, Bosnia and Herzegovina  
e-mail: [jasmina.selimovic@efsa.unsa.ba](mailto:jasmina.selimovic@efsa.unsa.ba)

for the transferred risk, is one of the main tasks of actuaries in insurance companies that are involved in nonlife insurances. The ratemaking process for the portfolio of insured and the rate which would correctly distribute risk among the insured people is of key significance for insurers' business. On the one hand, the insurance premium should be economically acceptable for the insured who will pay it while on the other, from the insurance company's perspective, the premium should be high enough to ensure the adequate and timely payment of claims.

In the process of nonlife insurance ratemaking, risk classification is one of the most important elements. Classification ratemaking is a procedure of grouping insurance policies of a given insurance portfolio into homogenous groups with similar expected claims experience or risk profile so that all the insureds in the same group pay the same insurance premium (Werner and Modlin, 2010). Since the insureds of a particular risk class typically have similar features, insurers can better estimate the insurance premium. The general goal of risk classification is to determine a fair premium for each insured in the portfolio, i.e. to indirectly ensure the financial stability of the insurance company.

Risk classification is very important in the competitive insurance market, where price liberalization is in effect. A better understanding of the actual impact of risk classes on insurance ratemaking can help insurance companies to improve their financial position after the deregulation of the insurance market. Since the deregulation of the motor hull insurance market in Bosnia and Herzegovina was initially scheduled for 2020 but was postponed till October 2022 and 2023 (in the two Entities in Bosnia and Herzegovina), we believe that the development and application of statistical models for insurance premium ratemaking based on risk classification is a current topic for our insurance market as well. It should be noted that insurance companies strive to strengthen their market position by introducing innovative solutions. A good example is data mining as a process that uses a variety of methods to discover patterns and relationships between data to make valid predictions.

The research hypothesis states that adequate data mining methods used for risk factor selection of the nonlife insurance premiums will improve the ratemaking process significantly in terms of its quality compared to the standard risk estimation models.

The paper is organized as follows: After the introductory part, the following section provides a review of the relevant literature in the field along with the methodology applied, including the generalized linear models and the applied data mining models (forward stepwise and neural networks), as well as the explanation of the primary data used in the research. The obtained results are discussed in the penultimate section, while the final section concludes the paper.

## 2 Generalized Linear Models for Premium Ratemaking: Conceptual Framework and Literature Review

The first illustration of the application of generalized linear models (GLMs) for the insurance premium ratemaking was provided by McCullagh & Nelder in their book (1989) on the example of the estimate of the average claims in the motor hull insurance. In the second half of the 1990s, the use of GLM for the nonlife insurance premium ratemaking expanded in the insurance industry, as a response to the demands of market deregulation in the EU countries. From that period on, GLMs have become a standard in the insurance industry for the nonlife insurance premium ratemaking. This chapter is based on McCullagh & Nelder (1989), who are the leaders of the GLM approach as the main statistical tool, nonlife insurance premiums ratemaking. Haberman and Renshaw (1996) and Renshaw (1994) showed how GLM can be used to analyze the claim frequency and severity.

Although actuaries are thought to have fully mastered GLM, improvements and enhancements of GLM for various applications in the insurance industry are still a hot topic (Jong and Heller, 2013; Hilbe, 2014; Frees and Lee, 2016; Garrido et al., 2016; Coskun, 2016).

Generalized linear models (GLMs) can be viewed as an extension of linear regression models on the family of exponential distributions with a special link function. GLMs extend the linear regression model in two directions:

- Distribution of probability. The linear regression model with the normal random response variable and constant variance is not suitable for the adequate insurance premium ratemaking.
- Expectation. In linear models, the expectation is a linear covariate function, while covariance is constant. In GLM, expectation transformation is a linear covariate function, while variance depends on the expected distribution.

GLMs are aimed at estimating the dependent response variable, (Y), based on a given number of independent variables  $X_i, i = 1, 2, \dots, n$  that we have information on. To determine the nonlife insurance premium, the variable Y can be one of the following variables:

- Number of claims = number of claims per risk exposure unit
- Claims severity = size of claims per harmful event
- Risk premium = claim severity per risk exposure unit
- Claims ratio = claim severity compared to insurance premium.

Independent variables  $X_i$  are called predictors or control variables. Potential predictors include the characteristics of a policy of the insured or characteristics of the insured subject matter (e.g. insured amount, insured's age, place of insurance, etc.) that affect the response variable.

Predictors in GLM can be categorical variables and continuous variables. Categorical variables or factors are variables the values of which indicate affiliation to one of several possible categories (e.g. gender, type of vehicle). Categorical variables can be numeric or non-numeric. An individual value that a categorical variable can

assume is called a “level”. Continuous variables or covariates are numeric variables the values of which can be associated with a particular interval of real numbers (e.g. insured’s age, insured amount). The aim of GLM is to transfer as much variability from a random component as possible to a systematic component, i.e. to explain most part of the outcome variability of response variable using predictors. GLM consists of three components:

- (GLM1) Random component: Distribution of the response variable  $Y_i$  for  $i = 1, 2, \dots, n$  is independent and belongs to the exponential family of distributions. In the paper by Nelder & Wedderburn (1972), distribution  $Y$  is a member of the exponential family of distributions and includes distributions such as normal, Poisson, gamma, inverse Gaussian, binomial, exponential, and other distributions. In later papers, GLM was expanded to the multivariate exponential family of distributions and some nonexponential families, such as the negative binomial distribution.
- (GLM2) Systematic component: Linear predictor is a linear function of independent predictors. Predictors are combined to give the linear predictor  $\eta$ :

$$\eta = X\beta$$

- (GLM3) Link function: The relationship between the random and systematic components is defined through a link function  $g$ , which is a differentiable and monotonous function to which the following applies:

$$E[Y] \equiv \mu = g^{-1}(\eta)$$

For modeling claim frequency or the number of claims, the most commonly used distributions are Poisson distribution and negative binomial distribution. The first choice for modeling the frequency or the number of claims in the literature is GLM with Poisson distribution, according to Antonio & Valdez (2010), Dionne & Vanasse (1988,1992), Denuit & Lang (2004), and Flynn & Francis (2009). Poisson distribution has the variance equal to expectation. The function of variance in Poisson distribution is  $V(\mu) = \mu$ , i.e. variance linearly increases with expectation. When expectation increases, the curve of Poisson distribution becomes more symmetrical and resembles a normal distribution. In claim frequency, variance is typically greater than the expectation. If the Poisson distribution is used in this case, variance can be underestimated and thus statistical measures of the model can be wrong. Goldburd, Khare & Tevet (2016) suggest that the classical approach by McCulloch & Nelder (1989) be followed in this case, since they used the extended Poisson distribution instead of Poisson distribution.

Hilbe (2007) points out that the variance higher than expected in claim frequency can be the result of a series of factors, which cannot be corrected using the extended Poisson distribution or negative binomial distribution. The most frequent reason is a great number of policies without reported claims, i.e. policies where the number of claims is 0 compared to the estimate of the number of claims obtained using Poisson or

negative binomial distribution. A great number of policies without reported claims occurs when an insured does not report, to the insurance company, claims up to the amount of the deductible to achieve a bonus on the insurance premium in the following insurance year. Yip & Yau (2005) illustrated the use of zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) distribution, when variance is greater than the expectation and when there is a surplus of zeros.

Distribution of the claim severity is more difficult to predict than claim frequency. Gamma and inverse Gaussian distributions are best suited for modeling claim severity or the average claim due to the positive values of the response variable. Gamma distribution is the natural choice and the most represented distribution for modeling the claim severity in GLM, as can be seen in Ohlsson & Johansson (2010), Parodi (2014), Kaas et al. (2009), and Pinquet (1997). One of the reasons for which gamma distribution is so appreciated is that the standard deviation is proportionate to expectation. The function of gamma distribution variance is  $V(\mu_i) = \mu_i^2$ , i.e. the variance of claim severity in the gamma model is proportionate to the exponential function of expectation. The function of the gamma distribution variance assigns a greater variance to claims that have a greater expectation, which is a desirable feature when claim severity is modeled in GLM, even if the scale parameter  $\theta$  is constant for all claims. Gamma distribution has a positively asymmetrical curve with a sharp peak and long right-ward tail which approaches 0. Like in gamma distribution, the function of the variance of inverse Gaussian distribution increases exponentially with expectation  $V(\mu_i) = \mu_i^3$ . The curve of inverse Gaussian distribution is positively asymmetrical with a sharper peak and wider tail and is employed in situations when greater curvature of distribution is needed.

Tweedie distribution has the best properties for modeling risk premium or claims ratio. Although modeling claims frequency and claim severity in GLM can provide a better understanding of the ways in which risk factors affect insurance premium, the direct application of Tweedie distribution for the risk premium ratemaking can often yield very similar results as those obtained using the combination of the model of claims frequency and the model of the average claim severity separately. Smyth & Jørgensen (2002) and Jørgensen & Souza (1994) discussed the use of Tweedie distribution for modeling the total claim severity. Besides the standard parameters of the exponential family  $\mu$  and  $\phi$ , Tweedie distribution has the third parameter  $p$ , called the *power parameter*, where  $p$  can assume any value except those between 0 and 1.

The linear predictor  $\eta_i$  is a linear function of independent predictors  $X_{ij}$  and unknown parameters denoted as  $\beta_0, \dots, \beta_p$  :

$$\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p + \xi_i, \quad i = 1, \dots, n$$

The value of the average is denoted by  $\beta$  and the values of parameters  $\beta_1, \dots, \beta_p$  are determined using the maximum likelihood method. The degree of freedom is defined as a difference between the number of data and the number of parameters. Besides the basic level in GLM, one parameter in the linear predictor is determined

for each level of the categorical predictor. For each continuous prediction variable, only one parameter in the linear predictor is determined.

Using the Log link function, the sum of linear predictor components is transformed into the product of linear predictor components; in other words, the additive model is transformed into a multiplicative model.

The multiplicative model is a widely used model for determining the premium, due to its advantages (it is simpler than additive models and more practical to use; with multiplicative models, the premium is always positive, without adjustments such as the introduction of the minimum premium; multiplicative models are more intuitive than additive models for expressing the impact of risk increase or decrease on the insurance premium).

For those reasons, the Log link function that yields multiplicative models is the best link function for the model of claim frequency and claim severity, although it does not necessarily have to be the natural link function. If the Log link function is used, it is better to use the Log of this variable rather than the continuous predictor. Inclusion of the Log continuous predictor into GLM with the Log link function provides flexibility in determining the curve of the response variable. In general, Log continuous predictors are used for Log link functions, while the original continuous predictors are used only in specific cases, e.g. when the predictor includes the value of 0 or when it presents the time or the trend.

An important property of GLM is the possibility to introduce offset in the definition of the linear predictor. In cases when the effect of the predictor on the response variable is known, it is appropriate to include information on this variable in the model as the offset, rather than the estimate of parameter  $\beta$  for this variable. Offset is defined as a predictor, the parameter of which equals 1.

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + offset$$

When an offset is included in GLM, it is crucial that it is on the same “scale” as the linear predictor. In the case of the Log link function, the Log of the offset variable should be determined before its inclusion in the model.

After the selection of the suitable GLM model in terms of the distribution function, covariate  $X_i$ , link function  $g(x)$ , offset  $\xi$ , scale parameter  $\phi$ , and weight  $\omega$  for the given set of data of the response variable  $\bar{Y}$ , it is necessary to estimate the unknown parameters  $\bar{B}$ . Parameters  $\beta_1, \dots, \beta_p$  are estimated by means of the maximum likelihood (ML) method.

Insurance premium modeling includes several thousand insurers’ data and therefore solving the system of equations uses iterative numeric methods: Newton–Raphson iteration method, Fisher iteration method, and iteration method of least squares. In practice, the most commonly used method is Newton–Raphson iteration method, whose formula goes as follows:

$$\beta_{n+1} = \beta_n - \bar{H}^{-1} \cdot \bar{s}$$

where  $\beta_n$  is the  $n$ th derivation of the vector of estimate parameter  $\bar{\beta}$  with  $p$  elements,  $\vec{s}$  is the vector of the first derivation of log-likelihood and  $\vec{H}$  is a  $p \times p$  matrix which includes the second derivation of log-likelihood. The iterative process begins with the estimation of parameter  $\beta_0$ , which equals 0 or is determined based on the previously defined estimation of parameter  $\beta_0$ .

Nelder & Wedderburn (1972) proved that the combination of Newton–Raphson and Fisher methods is asymptotically equivalent to the least-squares method since the distribution of parameters becomes identical when  $n \rightarrow \infty$ .

One of the fundamental advantages of GLM is the possibility to determine the reliability of the estimation of parameter  $\bar{\beta}$ . For each predictor included in GLM, one obtains the estimate of parameter  $\bar{\beta}$ . One of the crucial questions that arise is whether each predictor affects the response variable, i.e. whether it should be included in the model. GLM will benefit from including the predictors that systematically affect the response variable, as opposed to the predictors that have a random effect on the response variable. To establish the significance of a predictor for the model, a series of criteria can be taken into account.

Model validation consists of checking the model adjustment to data, i.e. its predicting performances. The process of model adjustment to data can be considered as a way of replacing the actual values of the response variable  $Y_i$  with the estimated values of a response variable  $\hat{Y}_i$  obtained from GLM. It is believed that a model is well adjusted to data if the estimated values of the response variable are close to the actual values.

The saturated model is defined as a model where the number of parameters equals the number of the response variable data so that the estimated values equal the actual values of the response variable. The saturated model is theoretically the best model since it perfectly predicts all data, though it is very complex at the same time.

The zero model is defined as a model which has a single estimate, i.e. a single parameter for all data, which is the average expected value. The zero model is simple; however, it does not represent the adequate structure of data.

The optimum model is one in between these two extremes. A simple model that well describes data has priority in practice compared to a more complex model which perfectly describes data. Therefore, the aim is to select the optimum model which will best explain the response variable  $Y_i$ , and has as few predictors as possible. The quality of model adjustment to training data is typically estimated by means of statistical measures: deviation and residuals.

If the model includes predictors which are a subset of a larger model's predictors, the smaller model is called the nested model. The nested model has the same data distribution function and the link function, though the linear predictor has fewer parameters compared to the larger model.

Nested models are compared when we consider including or excluding predictors in or from the model. One should also take into account the fact that adding predictors to the model always decreases deviation, regardless of whether the predictor affects the response variable or not. The inclusion of a larger number of predictors implies



a larger number of parameters, which provides the model with a higher degree of freedom in adjustment to data.

There are two distinct ways of selecting predictors: type I analysis and type III analysis. Type I analysis is used for considering the significance of the predictors which are added sequentially to the model. The starting point is the zero model as the basic one, and in each step, predictors are included and their significance tested. Type I analysis depends on the sequence of including predictors into the model. Type III analysis consists of testing the significance of all predictors included in the model. Type III analysis does not depend on the sequence of including the variables into the model, as is the case in type I analysis. Type III analysis is performed with predictors rather than the levels of categorical predictors.

If one wants to compare models which are not nested, the mere measure of deviation does not suffice. An increase in the number of parameters decreases deviation and consequently adding more parameters than needed can result in an excessive model adjustment to data. There are several statistical measures that can be employed for comparing different models and making decisions on the final model.

A practical way of testing the complexity of the model is the use of information criteria. Information criteria refer to the ratio of the accuracy of the model's adjustment to data to the model's complexity. The two most commonly used information criteria are:

- (1) Akaike information criterion (AIC), developed by Akaike (1974). AIC statistics has the following form:  $AIC = -2l(y_i, \hat{\mu}_i) + 2p$
- (2) Bayesian information criterion (BIC) developed by Schwarz (1978). BIC statistics is a measure similar to AIC and is defined as follows:  $BIC = -2l(y_i, \hat{\mu}_i) + p \cdot \ln(p)$

AIC and BIC information criteria are found in most standard statistical software and are widely employed in empirical studies. The information criteria allow the comparison of two models with different numbers of parameters. The lower the information criterion is, the model is considered to be better. Models with low deviation and high AIC or BIC should be dismissed. In a practical sense, AIC tends to yield "more reasonable" results. Over-reliance on BIC can result in excluding predictors from a model as described by Kuha (2004).

Gini coefficient or Gini index is typically used for measuring the inequality of national income.

### 3 Data Mining Methods: Conceptual Framework and Literature Review

Data mining is a process of discovering interesting patterns and knowledge out of a large amount of data. Data mining methods for searching for hidden patterns in data are generally classified into two categories: descriptive and predictive methods. Descriptive methods focus on the interpretation and understanding of relations among

data. Predictive methods are aimed at constructing models that will predict one or more variables based on the data.

The comprehensive overview of data mining methods presented by the authors Guo (2003), Han, J. et al. (2012) and Hastie et al. (2001) systematically presented most of the statistical methods used in data mining today. Sumathi and Sivanandam (2006) explored the concepts of data mining and data warehousing and presented areas of application in the insurance industry. Francis (2001) compared neural networks and regression models on insurance examples. Dugas et al. (2003) investigated the application of neural networks to determine motor insurance premiums in North America. Shapiro and Jain (2003) presented a collection of papers by various authors on the theory and application of data mining methods in the insurance industry. Yao (2008) used cluster analysis methods to determine the claims frequency by geographical areas. The work of Kolyshkina et al. (2004) discussed the advantage of combining GLM with a multivariate adaptive regression mining method. Lowe and Pryor (1996) compared the methods of neural networks and GLM and concluded that neural networks have a more general application than GLM and suggested certain possibilities of using neural networks in insurance but concluded that computationally demanding neural networks can prevent their wider application in insurance.

The insurance industry uses different data mining methods, from classification, cluster analysis to regression. This research employs data mining methods that can be applied for developing predictive models, based on stepwise regression and neural networks.

### ***3.1 Stepwise Regression***

Stepwise regression is a method of regression analysis that selects predictors by means of an automatic procedure. Automatic procedures of variable selection are useful when there are many independent predictors for which a subset of variables that will be the subject of further modeling should be determined. Such procedures are convenient in the stage of preparing data for modeling when predictors that should be included in the model cannot be determined based on theory or practice.

Stepwise regression is one of the most commonly used methods for the selection of predictors. As its very name suggests, this procedure adds or removes independent predictors in each step based on the defined criteria. A widely applied algorithm of stepwise regression was first proposed by Efroymson (1960). The three most commonly used procedures of predictor selection in stepwise regression are forward selection, backward elimination, and forward stepwise regression.

The forward selection procedure starts with the zero model. Then, the predictor that has the greatest correlation coefficient with the criterion variable is included first. In each following step, the model is supplemented with the predictor that has

the greatest coefficient of partial correlation with the criterion variable. The procedure stops when there are no more statistically significant predictors that affect the response variable and that should be added to the model.

The backward elimination procedure starts with the model that includes all predictors. The predictor that has the smallest coefficient of correlation with the criterion variable is eliminated from the model. In each following step, the variable that has the smallest coefficient of partial correlation with the criterion variable is eliminated from the model. The procedure continues as long as there are predictors that satisfy all conditions for elimination.

Forward stepwise regression is a procedure similar to the forward selection one, except that predictors are eliminated from the model if they become insignificant when other predictors are added. This stepwise regression procedure of variable selection starts without predictors in the model and adds and eliminates predictors, one by one, until variables can no longer be added or eliminated according to the defined criteria. This method eliminates the least significant predictor and includes the most significant predictor in the regression model alternately in each step. Values that must be in line with the criterion variable are calculated in each step. Predictors are typically correlated, and it cannot be predicted in advance how the inclusion or elimination of some of the predictors will affect the statistical significance of other predictors in the model.

The stepwise regression method uses the following criteria for adding and eliminating variables in the model: F-statistics, maximum adjusted value R2, minimum Akaike information criterion (AIC), and mean squared error (MSE).

There are slight differences in the procedure of variable selection, depending on the criterion of model selection. The traditional criterion for adding and eliminating predictors is based on F-statistics and corresponding p-values, which are compared to the defined values of significance for the inclusion or elimination of predictors. In the forward selection and backward elimination procedures, the F-statistics criterion typically uses p-values of 0.05 or 0.10. The forward stepwise regression procedure considers a variable to be statistically significant if the p-value in including the variable in the model is lower than 0.05 and the p-value in eliminating the variable from the model is higher than 0.10. The remaining three criteria compare the statistics (adjusted R2, AIC, or ASE) of the model after adding or excluding predictors to the current model. The procedure of variable selection stops when the optimum value is achieved (maximum value for criterion R2 and minimum value for AIC and ASE).

This paper will use the following regression methods for the risk factor selection:

- (1) Forward stepwise selection based on F-statistics (stepwise regression)
- (2) Forward stepwise selection based on AIC criterion (stepwise AIC regression).

F-statistics for adding or excluding variables from the current model are as follows:

$$F_{enter_j} = \frac{(SS_{e_p} - SS_{e_{p+1}})/l^*}{SS_{e_p}/(N - p^r)} \quad \text{and} \quad F_{remove_j} = \frac{(SS_{e_{p-1}} - SS_{e_p})}{SS_{e_p}/(N - p^c)}$$

and their corresponding p-values are as follows:

$$p_{enter_j} = P(F_{l^*, N-p^r} \geq F_{enter_j}) \text{ and } p_{remove_j} = P(F_{l^*, N-p^c} \geq F_{remove_j})$$

An AIC criterion for including and excluding a variable from the current model uses the following value:

$$AIC = N \ln \left( \frac{(N - 1) S_{yy} \times \tilde{r}_{yy}}{N} \right) + \frac{2p^r N}{N - p^r - 1}$$

The most common criticism of the application of the stepwise regression method is based on the belief that standard statistical tests are not suited for use in each step of the selection of predictors (Harrell, 2001). Standard errors of parameter estimates are considered underestimated, which results in excessive model adjustment. Despite the criticism, this method never ceased to be used and its application was revived for the needs of data mining in cases with a large number of potential predictors. Indeed, Famoye & Rothe (2003) established that the application of stepwise regression is acceptable in practice if it is performed with an appropriate degree of caution.

### 3.2 Neural Networks

Artificial neural networks (ANN) are software or hardware sets that use the iterative procedure from previous data to attempt to find a link between input and output model variables, to obtain the output value for new input variables (or, in other words, value learned by examples). Further in the text, the term “neural networks” will refer to artificial neural networks.

The development of neural networks began in the 1950s. Hebb (1949) formulated the rule of neural networks learning, which is considered the first important contribution to the development of the theory of neural networks.

Although the concept of artificial neural networks was introduced in the mid-twentieth century, they became popular only with the development of databases and better-performing computers. Neural networks were criticized due to the harder interpretation of results and a lack of diagnostic measures, which initially made neural networks a less desirable method for data mining. Advantages of neural networks include a high degree of tolerance of errors in data, and the ability to classify data even in situations when there is little knowledge of the relations between predictors and response variables. Neural networks are convenient for continuous predictor and response variables as opposed to most other algorithms. These are the factors that contribute to the usefulness of neural networks for solving problems of classification and predictions in data mining. The development and study of artificial neural networks are based on the existing knowledge of the way of human brain’s functioning. Traditionally, the concept of neural networks refers to the biological neural network. Artificial neuron functions in a way similar to a biological neuron, i.e. it receives some input data (dendrites) and, after processing, forwards the data (axon).

Like the biological neuron, the artificial neuron has several inputs it receives information from. Inputs in neurons, indexed with  $i = 1, \dots, n$  receive input values  $x_i$ . Input values are real numbers. Each input value  $x_i$  is multiplied by the weight value  $w_i$ . The sum  $S$  of all the weighted values is called internal activation  $I$ . It equals: 
$$I = S = \sum_i w_i * x_i.$$

Thus obtained sum  $S$  (or internal activation  $I$ ) is processed by means of the activation function  $f$ . If  $S$  has a value that is above the threshold of the observed neuron, the neuron will be efficient and will modify the signal with its activation function  $f$ . Output from the neural network equals  $y = f(I)$ .

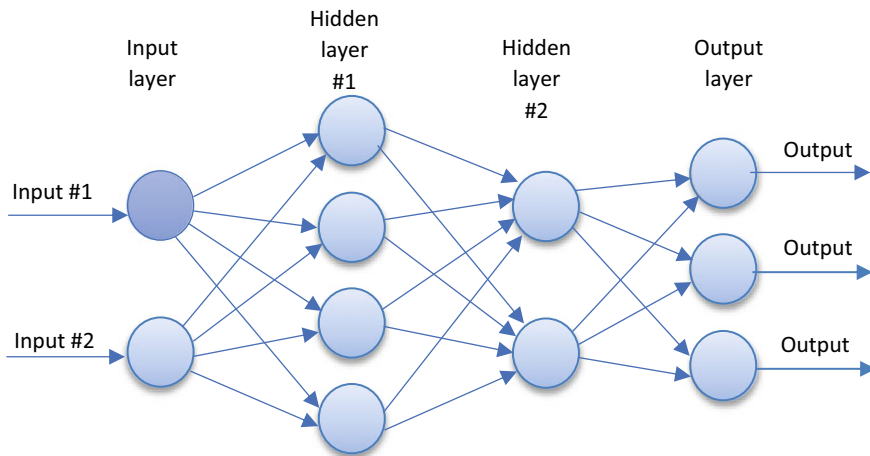
Function  $f$  symbolizes the intelligent function of the brain. Biological and artificial neurons have almost the same conceptual composition, and it is only the ways of performing individual functions that differ. Neurons in an artificial neural network are organized into layers where information is processed in parallel.

Neural networks consist of three layers: input, output, and hidden (Kriesel, 2007). An example of a neural network that consists of the input, output, and two hidden layers is presented in Fig. 1.

Input and output layers are those which lead data into or take them out of the network, while the hidden layer is an inter-layer that performs the desired function. A hidden layer can have an arbitrary number of sub-layers, depending on the needs and complexity of the given network. It is the hidden layer where inter-dependences in the model are learned. Information is processed in the hidden layer and sent to the neurons of the output layer.

There are two basic types of neural network structure:

- Feedforward neural networks allow the signal to move only in one direction, from input to output. Each neuron in a layer affects only neurons in the next layer, which is closer to output. These networks give the result very fast.



**Fig. 1** Example of the structure of neural network

- Feed-back neural networks allow signals to move in both directions because nodes are inserted in the network. In feed-back neural networks, input and output neurons are not explicitly defined. Such structures of neural network are used for solving dynamic problems, though they can become complicated and require a lot of time for adjusting.

The process of neural network design consists of several stages. The following stages are typically used in the literature:

- (1) Definition of the model; it includes the selection of input and output variables, collection, and preparation of input data where the neural network will be applied. The collected data are classified into two data subsets: training data and test data.
- (2) Selection of neural network algorithm; supervised and unsupervised learning.
- (3) Determination of network architecture; it pertains to the way of linking individual neurons in the entire network.
- (4) Determination of activation functions between layers; defining mutual links between input and output of the neural network.
- (5) Selection of learning rules and learning parameters; a formula that is used for adjusting the weights of links between neurons.
- (6) Selection of measures for assessing the neural network; it is defined through the prediction error, which is a difference between the actual and the predicted value.
- (7) Construction of neural network; when the model has been defined, input data prepared, and the neural network algorithm, as well as learning rule and necessary functions selected, the network should be taught or trained on the prepared data, so it can recognize the correlation between the data and be able to predict outputs based on input values.

Many algorithms have been developed for neural networks; however, the error backpropagation algorithm has so far had the widest commercial use. This algorithm has been decisive for the widespread application and popularity of neural networks in different areas. The error backpropagation algorithm has the structure of a multilayer feedforward neural network. It is a universal algorithm of network learning applicable to prediction problems, where the value of one or more output variables is predicted, as well as to classification problems, where input variables are distributed into one of the classes of response variable defined at the output. The network structure consists of the input layer, one or more hidden layers, and the output layer.

This paper uses the error backpropagation algorithm that consists of the input layer, one hidden layer, and the output layer. A neuron calculates its output, based on the sum of weight factors of its inputs based on the sigmoid activation function. For neuron  $j$ , this procedure is illustrated by the following equations:

$$I_j = \theta_j + \sum_{i=1}^n w_{ji}x_i$$

$$y_j = f(I_j)$$

Data mining methods can be used to overcome the problems of traditional GLMs, and to improve the performance of the risk premium predictive model. Some of the authors combined data mining methods and GLM to take advantage of both approaches. Kolyshkin et al. (2004) discussed the advantages of combining GLM with the multivariate adaptive regression splines (MARS) method. Also, they have compared different data mining methods (stepwise regression, decision tree, etc.) for the selection of predictors on the example of a household property insurance premium. Recent papers dedicated to GLM, by Makov and Weiss (2016), Coskun (2016) include stepwise regression in variable selection. Francis (2001) compared neural networks and regression models on insurance examples and used neural networks to select risk factors. The works of these authors were an incentive to explore data mining methods for the selection of variables in the data preparation phase, in order to improve the predictive performance and efficiency of the GLM risk premium predictive model.

## 4 Empirical Data and Analysis

The research was conducted on the motor hull insurance dataset of one of the leading insurance companies in Bosnia and Herzegovina. Having in mind that the motor hull insurance product has the same insurance coverage in all companies in BiH, and that the insurance company whose data were used in the survey has a significant share in the insurance market, we believe that the data used are a representative sample for this survey. The insurers' data for the last five successive years are a good basis for the model development. The sample consists of 18,012 policies of the insureds' passenger vehicle insurance and the data on claims history.

After the exceptions were excluded, the dataset contains 22 variables and consists of 17,404 records on motor hull insurance policies during five consecutive years. The dataset was divided by random distribution: 80% for training and 20% for testing and model evaluation.

The research was conducted with the CRISP-DM methodology using the IBM SPSS Modeller software package.

### 4.1 Selection of Modeling Techniques

The risk premium approach is traditionally used for nonlife insurance premium ratemaking, and it will be therefore employed in this research as well. Risk premium refers to the expected amount of all claims reported by the insured during the insurance period, and is obtained by multiplying two components, the expected value of

claims frequency and the expected value of the average claim severity:

$$E \left[ \sum_{i=1}^N C_i \right] = E[Y] \times E[C_i]$$

for claim severity ( $C_1, C_2, \dots$ ) regardless of the number of claims ( $Y$ ). Naturally, it applies only under the assumption that the expected values of claim frequency and the expected values of the average claim severity are independent of each other.

Having in mind that different risk factors affect the frequency of claims and the average claim severity, the risk model for determining the risk premium has been developed based on two models:

- GLM for estimating claim frequency
- GLM for estimating claim severity

The standard GLM for estimating claim frequency was developed subsequently and was used as the reference for comparing to predictive models, which include the data mining methods.

Furthermore, different data mining methods for risk factor selection were considered. The following data mining methods for the selection of predictors were used for the risk factor selection: stepwise regression and neural networks. Models were created for each method for variable selection, which result in a smaller set of predictors. The models were assessed and the best data mining method for the risk factor selection was selected.

It was followed by the risk class selection for multiple categorical predictors and continuous predictors using the data mining methods. Models were developed for the described data mining methods and, as a result, new predictors with a smaller number of categories, i.e. risk classes were obtained. The new predictors, obtained based on the described data mining methods, were used as input variables for GLM for claims frequency. Based on the model assessment, the best data mining method for the risk class selection was selected.

## 4.2 *GLM for Claim Frequency Estimate*

The standard approach to developing the GLM model for estimating claim frequency in the way that is customary in the actuarial practice is designed in the following five steps:

- (1) Model parameters were defined: function of distribution, link function, response variables, and predictors
- (2) Significance of each predictor was tested, as well as the significance of the interaction between predictors
- (3) A model was formed based on the significant predictors
- (4) Significance of each estimate parameter was tested



(5) The final GLM model was formed.

GLMs for claim frequency estimate the number of claims for each risk class over a given period of risk exposure. The distribution used the extended Poisson distribution as the most popular distribution for modeling claim frequency in nonlife insurance. Although GLM with Poisson distribution can also be applied to the continuous response variables, the number of claims was used for the response variable rather than claim frequency as the continuous variable for the sake of model simplicity. Since all insurance premiums in the dataset do not have the same risk exposure,  $\text{Log}(\text{exposure})$  was included in the model in calculating the number of claims as to the offset. It should be noted that both the number of claims and claim frequency yield the same estimated parameters and model statistics. For the link function, the natural link function,  $g(x) = \ln(x)$ , was used to make the model multiplicative. Scale parameter  $\phi$  for Poisson distribution equals 1.

All 14 categorical predictors were included in GLM to obtain a model with the best AIC information criterion. Predictors are presented with the original records of the training dataset at the level of the insurance policy. For nominal variables, the category with the greatest risk exposure was taken as the base risk class, while the smallest category was taken as the base risk class for ordinal variables.

Testing of the hypothesis on the significance of predictors used Wald test and type III analysis, to determine variables that should be kept in the model. Type III analysis tests each predictor, under the assumption that all the variables are included in the model. Wald test was used to assess the significance of each predictor, taking into account all the other predictors. Wald test follows the  $\chi^2$  distribution with the statistically significant value  $p \leq 0.001$  and df degrees of freedom, which represents the number of parameters associated with the analyzed variables.

The  $p$ -value of the Wald test determines the impact of each predictor on the response variable.

The variable denoted by  $\text{Osig\_suma}$ , which represents the insured amount, is not statistically significant, since the  $p$ -value of 0.628 is greater than the set level of significance  $p < 0.001$ , and this variable was therefore eliminated from the model. All the other predictors which had the significance level  $p < 0.001$  were also eliminated from the model (Table 1).

The significance of the interaction between two categorical variables:  $\text{MarkaD}$  and  $\text{KlasaD}$  was examined since it is justified for business reasons. It can be concluded from Table 2 that the inclusion of the interaction between these two variables is not statistically significant, and the interaction will consequently be not included in the final model.

After all predictors the  $p$ -value of which is greater than the statistically significant value ( $p < 0.001$ ) have been eliminated from the model, the remaining variables are statistically relevant, which clearly indicates their effect on claim frequency. The results of GLM presented in Table 3 confirm the importance of the following risk factors for claim frequency: make of the vehicle, class of the vehicle, the purpose of the vehicle, leasing, contract duration, and type of insurance.

**Table 1** Wald test for Poisson GLM: original predictors

Predictors	Wald $\chi^2$	df	p-value
Average	55.924	1	0.000
Age	18.276	16	0.308
MarkaD	73.988	28	0.000
KlasaD	40.898	6	0.000
NamjenaD	16.277	1	0.000
LeasingD	17.875	1	0.000
Trajanje_ugD	16.917	1	0.000
Tip_osigD	23.053	1	0.000
Vid_pribaveD	10.370	3	0.016
AO_polisaD	0.166	1	0.683
Lojalnost	2.153	2	0.341
Vel_OpcineD	1.328	1	0.249
OpcinaD	129.585	89	0.003
Osig_suma	0.235	1	0.628
Power	0.834	1	0.361

**Table 2** Wald test for Poisson GLM: interaction between predictors

Predictors	Wald $\chi^2$	df	p-value
Average	251.233	1	0.000
MarkaD	41.326	28	0.050
MarkaD * KlasaD	79.967	66	0.116
KlasaD	17.628	6	0.007

**Table 3** Wald test for Poisson GLM: standard approach to the risk factor selection

Predictors	Wald $\chi^2$	df	p-value
Average	195.535	1	0.000
MarkaD	84.281	28	0.000
KlasaD	41.391	6	0.000
NamjenaD	17.131	1	0.000
LeasingD	32.109	1	0.000
Trajanje_ugD	23.562	1	0.000
Tip_osigD	31.901	1	0.000

Measures of model adjustment to data for claim frequency are presented in Table 4. The estimate of scale parameter  $\phi$  obtained based on the total deviation 0.92 and Pearson's moment estimator 0.991 is smaller than 1, which shows that the variance is lower than expected, and it can be concluded that Poisson distribution is adequate for estimating claim frequency.

**Table 4** Measures of Poisson GLM adjustment: standard approach to the risk factor selection

Model adjustment measures	Value	df	Value/df
Deviation	12,809	13,927	0.920
Scaled deviation	12,809	13,927	
Pearson's moment estimator	13,807	13,927	0.991
Scaled Pearson's moment estimator	13,807	13,927	
Log-likelihood	-11,843		
Akaike information criterion (AIC)	23,764		
Corrected AIC (AICC)	23,764		
Bayes information criterion (BIC)	24,058		
Consistent AIC (CAIC)	24,097		

Although predictors MarkaD, KlasaD, NemjenaD, LeasingD, Trajanje\_ugD, and Tip\_osigD are statistically significant for claim frequency, it does not mean that each category of these variables is statistically significant. Results of GLM for claim frequency are satisfactory for risk factors NamjenaD, LeasingD, Trajanje\_ugD, and Tip\_osigD, since these are categorical variables with low cardinality, where a sufficient number of records is included for each category. The significance of each category for categorical predictors will be estimated using the horizontal line test. The horizontal line test is used to establish whether a function is injective. If the horizontal line intersects the function's graph several times, the function is not injective. In this specific case, if a horizontal line can be drawn between the categories without intersecting the graph of confidence function, it can be claimed that the variable satisfies the horizontal line test, i.e. that the categories are statistically significant.

The value of predictor LeasingD was estimated with the 95% confidence interval. The graph clearly shows that LeasingD satisfies the horizontal line test, i.e. that a horizontal line can be drawn in the confidence interval between categories 0 and 1. This leads to the conclusion that both categories of predictor LeasingD are statistically significant.

However, a problem arises with the predictors MarkaD and KlasaD, since these are categorical variables with a larger number of categories, for which GLM results show a high degree of uncertainty. MarkaD is a multiple categorical variable with 29 categories. MarkaD is a significant variable for claim frequency, and a large number of categories of this variable show a high standard error and a  $p$ -value higher than 0.001. The reason for this is the fact that a large number of categories of MarkaD do not have sufficient risk exposure, i.e. the sufficient number of insurance policies to be included in GLM as parameters.

The alternative solution to this problem is to group categories that do not have sufficient exposure. In this way "new" categories are created, i.e. risk classes with greater exposure that can yield credible results with GLM.

The standard approach in grouping categorical variables is adding the risk classes which are not statistically significant to the base risk class, to obtain the adjusted

**Table 5** Wald test for Poisson GLM: standard approach to risk class selection

Predictors	Wald $\chi^2$	Df	p-value
Average	34.149	1	0.000
NamjenaD	22.559	1	0.000
LeasingD	45.687	1	0.000
Trajanje_ugD	30.758	1	0.000
Tip_osigD	33.209	1	0.000
MarkaT	21.543	3	0.000
KlasaT	56.877	3	0.000

predictors with sufficient exposure. In this case, the original predictors MarkaD and KlasaD were replaced with new categorical variables MarkaT and KlasaT, which have a smaller number of categories. Type III analysis of the significant predictors on claim frequency with included new variables MarkaT and KlasaT is presented in Table 5.

The original categorical predictors: NamjenaD, LeasingD, Trajanje\_ugD, Tip\_osigD, and the new derived variables MarkaT and KlasaT were included in the Poisson GLM model.

Table 6 presents the model adjustment to data with grouped multiple categorical variables and reveals that AIC increased, due to a decrease in degrees of freedom. The estimate of the scale parameter  $\hat{\phi}$ , obtained based on the total deviation and Pearson’s moment estimator increased, though near the value of 1, which indicates that the variance is close to the expected value and that Poisson is still an adequate function of the distribution.

Parameter estimation, standard error, confidence interval, Wald test, and  $p$ -value for Poisson GLM based on the standard approach to the risk factor selection and risk classes are presented in Table 7. All parameters for the selected risk factors and created risk classes are statistically significant ( $p < 0.05$ ), which is satisfactory for the calculation of the expected values of claim frequency.

**Table 6** Measures of Poisson GLM adjustment: standard approach to risk class selection

Model adjustment measures	Value	df	Value/df
Deviation	12,886	13,955	0.923
Scaled deviation	12,886	13,955	
Pearson	13,990	13,955	1.003
Scaled Pearson Chi-Square	13,990	13,955	
Log-likelihood	-11,881		
Akaike information criterion (AIC)	23,784		
Corrected AIC (AICC)	23,784		
Bayes information criterion (BIC)	23,867		
Consistent AIC (CAIC)	23,878		

**Table 7** Estimation of Poisson GLM parameters: a standard approach to risk class selection

Parameter	$\beta$	SE	95% Wald CI		Hypothesis test		
			Lower limit	Upper limit	Wald $\chi^2$	df	p-value
Average	- 0.821	0.023	-0.867	-0.776	1,261.309	1	0.000
NamjenaD = 1	-0.160	0.033	- 0.226	-0.094	22.559	1	0.000
NamjenaD = 0	0						
LeasingD = 1	0.248	0.036	0.176	0.320	45.687	1	0.000
LeasingD = 0	0						
Trajanje_ugD = 1	-0.261	0.047	-0.353	-0.168	30.758	1	0.000
Trajanje_ugD = 0	0a						
Tip_osigD = 1	0.217	0.037	0.143	0.291	33.209	1	0.000
Tip_osigD = 0	0						
MarkaT = 4	0.739	0.258	0.232	1.246	8.154	1	0.004
MarkaT = 3	-0.980	0.448	-1.859	-0.101	4.778	1	0.029
MarkaT = 2	0.164	0.055	0.056	0.272	8.855	1	0.003
MarkaT = 1	0						
KlasaT = 4	0.277	0.062	0.156	0.399	19.989	1	0.000
KlasaT = 3	0.074	0.032	0.011	0.137	5.362	1	0.021
KlasaT = 2	-0.149	0.034	-0.216	-0.081	18.718	1	0.000
KlasaT = 1	0						

### 4.3 Development and Assessment of Models for Risk Factor Selection

This section discusses the data mining methods for selecting predictors out of the total number of predictors. The following methods are discussed for the variable selection method:

- Stepwise regression
- AIC stepwise regression
- Neural networks

The listed methods allow the estimation of predictors by their significance for the response variable. The variable significance indicates which predictors are important for a more accurate prediction of the response variable. In other words, they identify predictors that the data mining model uses the most to predict the response variable.

The importance of an individual predictor for the response variable does not say anything about the accuracy of the model itself. For neural networks, the importance of predictors is calculated using sensitivity analysis. Sensitivity measures the extent to which prediction error increases when one of the predictors is eliminated. More information on sensitivity analysis can be seen in Saltelli et al. (2004) and Francis (2001). Models for each of the listed methods were developed, with all the predictors included. The relative importance of each predictor in the model estimation is presented by means of a variable significance graph.

The methods were analyzed directly on the response variable: claim frequency. The significant predictors determined based on the selected data mining method were included in GLM for claim frequency.

The models were assessed, i.e. GLMs where risk factors were determined using the standard approach were compared with GLMs where risk factors were determined based on the previously described data mining methods. The criteria for ranking and selection of the best GLM, based on the approach to risk factor selection, are the following:

- (1) Model's adjustment to data and
- (2) Predictive performances of the model

All the models were developed on the training dataset, and comparisons of models' performances were conducted on the test dataset.

#### ***4.4 Risk Factor Selection for Claim Frequency***

The text below presents graphs of predictors' significance for claim frequency, using each of the data mining methods. Predictors significant for the response variable of claim frequency obtained based on the stepwise regression and AIC stepwise regression methods are presented in Figs 2 and 3, respectively.

Results of the estimates using the stepwise regression and AIC stepwise regression methods expectedly yielded similar results. Variables OpcinaD and MarkaD make up over 80% impact on claim frequency in both methods, whereby variable OpcinaD has over 70% impact on claim frequency. The significant predictors for claim frequency selected based on the neural network method can be seen in Fig. 4. The neural network model was created using the error backpropagation algorithm, with 14 predictors in the input layer and the response variable of claim frequency in the output layer. Results of the prediction by means of the neural network show that OpcinaD with 71% is by far the most significant predictor for claim frequency.

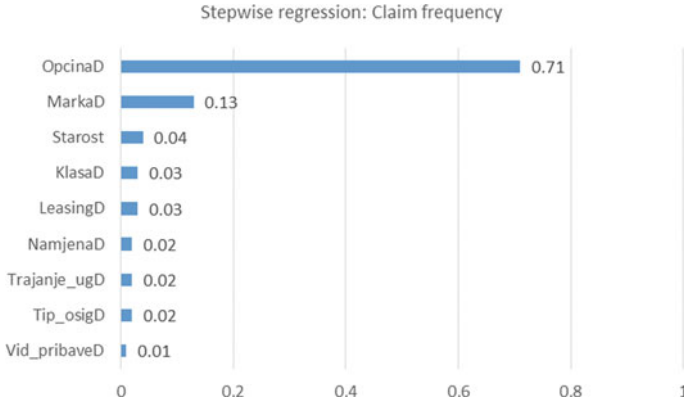


Fig. 2 Predictors for claim frequency based on stepwise regression

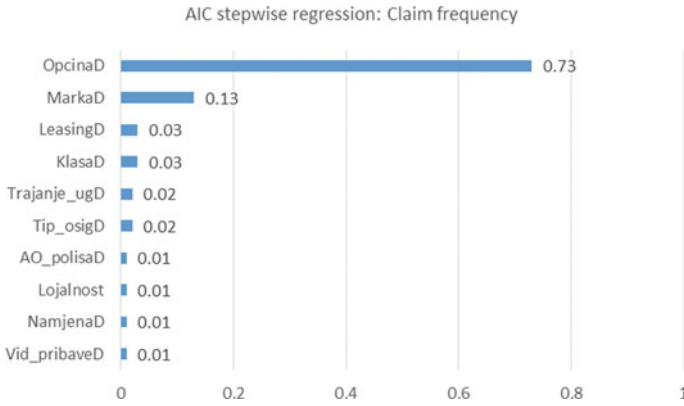


Fig. 3 Predictors for claim frequency based on AIC stepwise regression

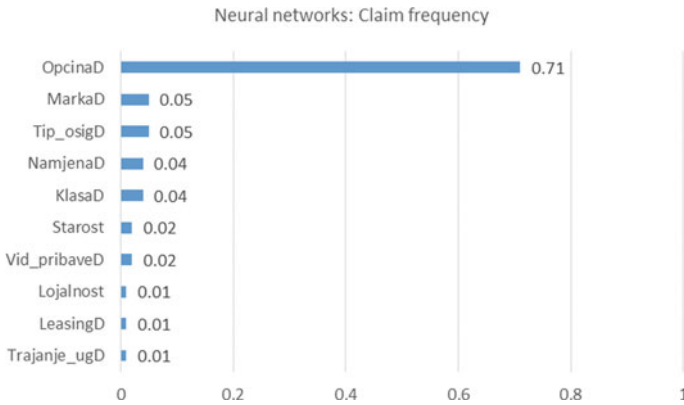


Fig. 4 Predictors for claim frequency based on neural network

### 4.5 GLM for Claim Frequency Estimate

The data mining method for variable selection yielded the optimum number of variables that were included in GLM Poisson with Log link function for the claim frequency estimate. Statistically significant predictors using type III analysis were taken for risk factors. The selected risk factors with GLM Poisson based on the standard approach and the approach using the data mining method are presented in Table 8.

GLM was compared to risk factors obtained using the data mining method, to establish whether the data mining methods for risk factor selection affect the improvement of predicting claim frequency. First, model adjustment to data was examined using the Akaike information criterion, and given in Table 9.

The AIC information criterion of the GLM Poisson method for claim frequency does not show crucial differences between different methods for risk factor selection. The reason is that all the methods single out a similar number of risk factors out of the total number of predictors. Stepwise regression method, AIC stepwise regression method, and neural networks method show a better result compared to other methods.

**Table 8** Risk factors for claim frequency

Variables	Standard approach	Stepwise regression	AIC stepwise regression	NN
Osig_suma				
Starost				
MarkaD	x	x	x	x
Klasa_D	x	x	x	x
SnagaD				
NamjenaD	x	x	x	x
LeasingD	x	x	x	x
Trajanje_ugD	x	x	x	x
Tip_osigD	x	x	x	x
Vid_pribaveD				
AO_polisaD				
Lojalnost				
Vel_OpcineD				
OpcinaD		x	x	x

**Table 9** Ranking GLM for claim frequency

Method for risk factor selection	AIC	Gini coefficient
Standard approach	23,764	0.112
Stepwise regression	23,798	0.139
AIC stepwise regression	23,798	0.139
Neural network	23,798	0.139





**Fig. 5** Ranking claim frequency models based on Gini coefficient

Gini coefficient was used for the comparison of models' predictive performances and the more objective measurement of models' results on the test dataset (Fig. 5).

It can be concluded from the above described that the selection of the optimum number of predictors using stepwise regression, AIC stepwise regression, and neural networks before inclusion in GLM improves the predictive performances of the model for developing claim frequency. Having in mind the simplicity of use and interpretation of results, as well as the speed of performance of both stepwise regression methods compared to neural networks, the former two methods have an advantage in risk factor selection.

Better predictive performances of GLM Poisson model with the risk factors selected using stepwise regression and AIC stepwise regression are due to the inclusion of a larger number of significant predictors compared to the standard approach. Indeed, compared to the standard approach, the claim frequency model with risk factors selected using the stepwise regression method includes an additional significant risk factor, *OpcinaD*. Type III analysis for claim frequency model Poisson GLM, where risk factors were determined using the stepwise regression method, as well as the model adjustment to data, are presented in Tables 10 and 11.

#### **4.6 Development and Assessment of Models for Risk Class Selection**

In the previous chapter, risk factors were selected using the data mining method and were included in GLM for claim frequency. The best results were achieved by the stepwise regression methods. Although the described data mining methods for risk factor selection improved GLM performances, there is still a problem with multiple categorical predictors, since some categories do not have enough data, or are similar with respect to the impact on the dependent response variable. The inclusion of all

**Table 10** Wald test for Poisson GLM: stepwise regression for risk factor selection

Predictors	Wald $\chi^2$	df	p-value
Average	163.968	1	0.000
MarkaD	88.842	28	0.000
KlasaD	45.570	6	0.000
NamjenaD	19.384	1	0.000
LeasingD	17.578	1	0.000
Trajanje_ugD	26.907	1	0.000
Tip_osigD	20.893	1	0.000
OpcinaD	540.358	91	0.000

**Table 11** Measures of Poisson GLM adjustment: stepwise regression for risk factor selection

Measures	Value	df	Value/df
Deviation	12,646	13,828	0.915
Scaled deviation	12,646	13,828	
Pearson’s moment estimator	13,342	13,828	0.965
Scaled Pearson’s moment estimator	13,342	13,828	
Log-likelihood	– 11,761		
Akaike information criterion (AIC)	23,798		
Corrected AIC (AICC)	23,801		
Bayes information criterion (BIC)	24,839		
Consistent AIC (CAIC)	24,977		

categories of multiple categorical variables can result in the excessive parameterization of the model so that the parameters in GLM correspond to noise in data rather than the actual patterns in data. Grouping of categories of multiple categorical predictors requires fast and efficient procedures.

Using the data mining methods, categories that do not differ significantly, taking into account their impact on the dependent response variables, can be combined. In this way, the total number of categories can be decreased. Data mining methods can also be used for the discretization of continuous variables. In predictive modeling, when records of a continuous variable are joined to the interval, a new categorical predictor is formed, which is further used in GLM.

The decrease of the cardinality of categorical variables and discretization of continuous variables form new predictors with a smaller number of categories, i.e. risk classes, which have a sufficient risk exposure in each class.

In this section, we discuss the data mining methods for determining risk classes of multiple categorical predictors and continuous predictors. Neural networks were considered as a data mining method.

Models for this method were developed, which included:

- Multiple categorical predictors: MarkaD, KlasaD, and OpcinaD and
- Continuous predictor: Osig\_suma

The new variables obtained based on the data mining method were included in Poisson GLM for claim frequency.

GLM was assessed and compared to risk classes based on the standard approach and GLM with risk classes determined using the data mining method.

Criteria for ranking and selecting the best GLM with risk classes obtained as the result of the applied data mining methods are as follows:

- (1) Model adjustment to data and
- (2) Model’s predictive performances

Models for risk class selection and GLM were developed on the training dataset, while comparisons of GLM performances were made on the test dataset.

### 4.7 Selection of Risk Classes for Claim Frequency

Having in mind that multiple categorical variables MarkaD, KlasaD, and OpcinaD affect only the claim frequency in predictive models for risk class selection, claim frequency was used for each of these variables as the response variable.

### 4.8 GLM for Claim Frequency Estimate

Measures for the comparison of the Poisson GLM model for claim frequency are presented in Table 12. Akaike information criterion (AIC) was used as a measure for model adjustment to data. Gini coefficient on the set of test data was used for the reliability of parameter estimation.

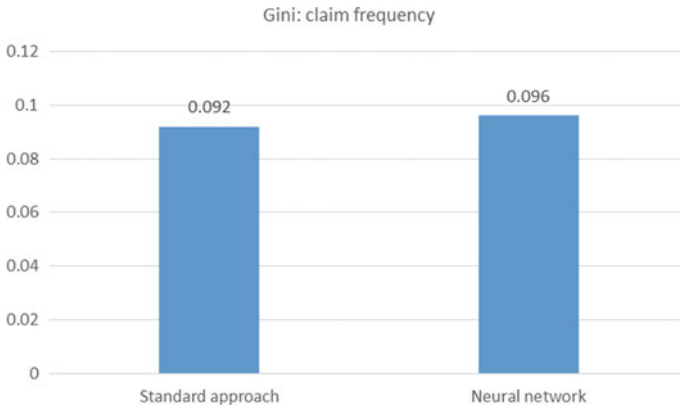
A lower AIC was observed in GLM for claim frequency with risk classes determined for multiple categorical predictors based on the neural network (Fig. 6).

The above discussion leads to the conclusion that the selection of optimum risk classes using the neural network method before inclusion in GLM improves the predictive performances of the model for claim frequency estimate.

The new predictors Marka\_CHAID, Klasa\_CHAID, together with the original predictors determined using the stepwise regression method were included in Poisson GLM.

**Table 12** Ranking GLM for claim frequency

Method of risk factor selection	AIC	Gini coefficient
Standard approach	23,784	0.092
Neural network	23,775	0.096



**Fig. 6** Ranking GLM for claim frequency based on Gini coefficient

**Table 13** Wald test for Poisson GLM: risk class selection with neural networks

Predictors	Wald $\chi^2$	df	p-value
Average	1,079.244	1	0.000
NamjenaD	19.361	1	0.000
LeasingD	26.300	1	0.000
Trajanje_ugD	30.452	1	0.000
Tip_osigD	25.731	1	0.000
Marka_CHAID	29.375	2	0.000
Klasa_CHAID	34.763	1	0.000
Opcina_CHAID	11.010	1	0.001

All the predictors are statistically significant and affect claim frequency, which can be seen from Table 13.

The measure for Poisson GLM model adjustment to data, based on risk classes determined using neural networks, is presented in Table 14.

Estimation of parameters from Table 15 reveals that each risk class created based on grouping categories using neural networks is statistically significant.

## 5 Conclusion

This research was aimed at examining the impacts of using data mining methods for risk factor selection on the nonlife insurance premium ratemaking, on the example of motor hull insurance. To this purpose, GLM for premium ratemaking and data mining methods for the selection of predictors were investigated for the selection of predictors.

**Table 14** Measures of Poisson GLM adjustment: risk class selection with neural networks

Measures	Value	df	Value/df
Deviation	12,885	13,955	0.923
Scaled deviation	12,885	13,955	
Pearson's moment estimator	13,855	13,955	0.993
Scaled Pearson's moment estimator	13,855	13,955	
Log-likelihood	- 11,879		
Akaike information criterion (AIC)	23,775		
Corrected AIC (AICC)	23,775		
Bayes information criterion (BIC)	23,843		
Consistent AIC (CAIC)	23,852		

Inclusion of all relevant risk factors that can affect claim frequency is of crucial importance for premium ratemaking, i.e. for calculating the insurance premium. To adequately select risk factors, this text took into account only some of the data mining methods. Particular attention was paid to the following methods: forward stepwise regression, forward AIC stepwise regression, and neural network.

Based on the obtained results of the Gini coefficient for claim frequency models, we can make the following conclusions:

- (1) Methods of stepwise regression, AIC stepwise regression, and neural networks achieve very good results for claim frequency.
- (2) GLM, with the application of any data mining method for risk factor selection, achieves better results in estimating claim frequency compared to the standard approach to risk factor selection.

Based on the obtained results of the Gini coefficient for the Poisson GLM model of claim frequency, it can be concluded that the application of neural networks as a data mining method for risk class selection achieves better results for claim frequency compared to the standard approach to variable selection.

The selection of the optimum number of predictors using stepwise regression, AIC stepwise regression, and neural networks before their inclusion in GLM improves predictive performances of the model for developing claim frequency. Having in mind the simplicity of use and interpretation of results, the speed of performing both stepwise regression methods compared to neural networks, the two described methods have the advantage in risk factor selection.

The use of the method for risk factor selection gives actuaries more time to refine the model while reducing the risk that some of the important risk factors have not been included in the model.

**Table 15** Estimation of Poisson GLM parameters: risk class selection with neural networks

Parameter	$\beta$	SE	95% Wald CI		Hypothesis test		
			Lower limit	Upper limit	Wald $\chi^2$	df	p-value
(Intercept)	-1.075	0.038	-1.149	-1.000	808.404	1	0.000
[NamjenaD = 1]	-0.148	0.034	-0.213	-0.082	19.361	1	0.000
[NamjenaD = 0]	0						
[LeasingD = 1]	0.192	0.037	0.118	0.265	26.300	1	0.000
[LeasingD = 0]	0						
[Trajanje_ugD = 1]	-0.261	0.047	-0.354	-0.168	30.452	1	0.000
[Trajanje_ugD = 0]	0						
[Tip_osigD = 1]	0.192	0.038	0.118	0.266	25.731	1	0.000
[Tip_osigD = 0]	0						
[Marka_NN = 3]	0.173	0.035	0.105	0.241	24.691	1	0.000
[Marka_NN = 2]	0.106	0.030	0.048	0.165	12.690	1	0.000
[Marka_NN = 1]	0						
[Klasa_NN = 2]	0.189	0.032	0.126	0.252	34.763	1	0.000
[Klasa_NN = 1]	0						
[Opcina_NN = 2]	0.098	0.029	0.040	0.156	11.010	1	0.001
[Opcina_NN = 1]	0						

## References

Antonio K, Valdez EA (2010) Statistical concepts of a priori and a posteriori risk classification. *Adv Stat Anal* 96(2):187–224

Coskun S (2016) Introducing credibility theory into GLMs for ratemaking on auto portfolio. Institute de Actuaries, Actuarial thesis. Centre d’Etudes Actuarielles

Denuit M, Lang S (2004) Non-life rate-making with Bayesian GAMs. *Insur: Math Econ* 35(3):627–647

Dionne G, Vanasse C (1988) A generalization of actuarial automobile insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bull* 19(2):199–212

- Dionne G, Vanasse C (1992) Automobile insurance ratemaking in the presence of asymmetrical information. *J Appl Economet* 7(2):149–165
- Dugas C, Bengio Y, Chapados N, Vincent P, Denoncourt G, Fournier C (2003) Statistical learning algorithms applied to automobile insurance ratemaking. *Casualty Actuar Soc Forum* 1(1):179–214
- Efroymson MA (1960) ‘Multiple regression analysis. In: *Mathematical methods for digital computers*’. Wiley, New York
- Famoye F, Rothe DE (2003) Variable selection for poisson regression model. *J Mod Appl Stat Methods* 2(2):380–388
- Flynn M, Francis LA (2009) More flexible GLMs: zero-inflated models and hybrid models. *Casualty Actuar Soc E-Forum* 148–224
- Francis L (2001) Neural networks demystified. *Casualty actuarial society forum*. pp 253–320
- Frees EW, Lee G (2016) Rating endorsements using generalized linear models casualty actuarial society. *Var Adv Sci Risk* 10(1):51–74
- Garrido J, Genest C, Schulz J (2016) Generalized linear models for dependent frequency and severity of insurance claims. *Insur: Math Econ* 70:205–215
- Goldburd M, Khare A, Tevet D (2016) Generalized linear models for insurance rating. *Casualty Actuar Soc* 5, 2nd edn
- Guo L (2003) Applying data mining techniques in property/casualty insurance. *Casualty Actuarial Society forum*. Available at: <https://www.casact.org/pubs/forum/03wforum/03wf001.pdf>
- Haberman S, Renshaw AE (1996) Generalized linear models and actuarial science. *Stat* 45(4):407–436
- Han J, Kamber M, Pei J (2012) *Data mining concepts and techniques*, 3rd edn. The Morgan Kaufmann, Burlington, USA
- Harrell F (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Chapter 5: resampling, validating, and simplifying the model. 3:88–103
- Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning*. Springer, New York, USA
- Hebb DO (1949) *The organization of behavior. A neuropsychological theory*. Wiley
- Hilbe J (2007) *Negative binomial regression*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511811852>
- Hilbe JM (2014) *Modeling count data*. Cambridge University Press, New York
- de Jong P, Heller GZ (2013) *Generalized linear models for insurance data*, 5th edn. Cambridge University Press, New York
- Jørgensen B, Souza M (1994) Fitting tweedie’s compound poisson model to insurance claims data. *Scandinavian Actuarial J* 1:69–93. <https://doi.org/10.1080/03461238.1994.10413930>
- Kaas R, Goovaerts M, Dhaene J, Denuit M (2009) *Modern actuarial risk theory, using R*. Springer, Berlin
- Kolyshkina I, Wong S, Lim S (2004) Enhancing generalised linear models with data mining. *Casualty actuarial society. Discussion paper program*. Available at [https://www.researchgate.net/publication/253447757\\_Enhancing\\_Generalised\\_Linear\\_Models\\_with\\_Data\\_Mining](https://www.researchgate.net/publication/253447757_Enhancing_Generalised_Linear_Models_with_Data_Mining)
- Kriesel D (2007) A brief introduction to neural networks
- Kuha J (2004) AIC and BIC: comparisons of assumptions and performance. *Socio Method Res* 33(2):188–229. <https://doi.org/10.1177/0049124103262065>
- Lowe J, Pryor L (1996) ‘Neural networks v. GLMs in pricing general insurance’, *General Insurance Convention*
- Makov UE, Weiss J, Frees EW, Meyers GG, Derrig RA (2016) Predictive modeling for usage-based auto insurance
- McCullagh P, Nelder JA (1989) *Generalized linear models*, 2nd edn. Chapman & Hall, London
- Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J Roy Stat Soc* 135(3):370–384
- Ohlsson E, Johansson B (2010) *Non-life insurance pricing with generalized linear models*. Springer-Verlag, Berlin

- Parodi P (2014) Pricing in general insurance, 1st edn. Chapman and Hall/CRC, New York
- Pinquet J (1997) Allowance for cost of claims in bonus-malus systems. *Ins: Math Econ* 19:152–152. [https://doi.org/10.1016/S0167-6687\(97\)81696-4](https://doi.org/10.1016/S0167-6687(97)81696-4)
- Renshaw AE (1994) Modeling the claims process in the presence of covariates. *ASTIN Bull* 24(2):265–285
- SAS Institute (2002) Data mining in the insurance industry - solving business problems using SAS enterprise miner software. Available at: <https://www.insurance-canada.ca/2002/10/01/data-mining-in-the-insurance-industry-solving-business-problems-using-sas-enterprise-miner-software/>
- Saltelli AS et al (2004) Sensitivity analysis in practice—a guide to assessing scientific models. Wiley
- Shapiro AF, Jain LC (2003) Intelligent and other computational techniques in insurance. *World Scientific*. <https://doi.org/10.1142/5441>
- Smyth G, Jorgensen B (2002) Fitting tweedie's compound poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin*. 32. <https://doi.org/10.2143/AST.32.1.1020>
- Sumathi S, Sivanandam SN (2006) Introduction to data mining and its applications. Springer-Verlag, Berlin
- Werner G, Modlin C (2010) Basic ratemaking. *Casualty actuarial society*, 4th edn
- Yao J (2008) Clustering in ratemaking: with application in territories clustering. *Casualty Actuar Soc Discuss Pap Program* 170–192
- Yip K, Yau, K (2005) On modelling claim frequency data in general insurance with extra zeros. *Ins: Math Econ* 36:153–163. <https://doi.org/10.1016/j.insmatheco.2004.11.002>



# Insurance Investments Management—An Example of a Small Transition Country



Željko Šain, Edin Taso, and Jasmina Selimović 

**Abstract** The COVID-19 pandemic during 2020 is the cause of the global health, economic, financial, and social crisis, which in terms of scale and possible harmful consequences has not been recorded since the global financial crisis in 2007–2008. For the BiH economy, risk exposure is particularly pronounced in terms of increased market risk of illiquidity of the economy, public and private companies, more precisely, illiquidity risk of (re) insurance company. The paper actualizes the application of modern methods in the management of the investment portfolio of a (re) insurance company, primarily methods for managing the risks of the securities portfolio and methods for managing credit risk. The starting point in this paper is the current domestic regulations on insurance, as well as the prescribed methodology for assessing the main risks and for quantitative analysis of the impact of the investment portfolio on the solvency of (re) insurance companies in BiH. To assess the risk of the securities portfolio, a model of variance and standard deviation was used, which show the degree of deviation of the security's return from the expected average return, while the return (positive, or negative, or zero) is measured by rising or falling or keeping the price of the security at the same level between two measurement periods. The relationship between yields between two securities is measured by the correlation coefficient and covariance.

**Keywords** COVID-19 pandemic · Crisis · Insurance industry · Investment risks · Securities risks · Credit portfolio

---

Ž. Šain · J. Selimović (✉)

School of Economics and Business, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

e-mail: [jasmina.selimovic@efsa.unsa.ba](mailto:jasmina.selimovic@efsa.unsa.ba)

Ž. Šain

e-mail: [zeljko.sain@efsa.unsa.ba](mailto:zeljko.sain@efsa.unsa.ba)

E. Taso

Insurance Supervisory Agency in FBiH, Sarajevo, Bosnia and Herzegovina

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

573

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research,*

*Data Science and Actuarial Studies*, Contributions to Economics,

[https://doi.org/10.1007/978-3-030-85254-2\\_34](https://doi.org/10.1007/978-3-030-85254-2_34)

## 1 Introduction

Ever since 2020, the COVID-19 pandemic has caused global health, economic, financial, and social crisis, the scale and possible harmful consequences of which have not been recorded since the 2007–2008 global financial crisis. This disease is a global threat to population's health and, at the same time, it causes mass interruptions in manufacturing and business and high unemployment, restricted movement and distribution, decrease of supply and demand, income and profits. Still, the development of the effective vaccine in the second half of 2020, and the started and partly effectuated vaccination throughout the world, which is planned to encompass about 75% of the world population, together with other, primarily healthcare measures, in all likelihood heralds the suppression of the pandemic. However, the disease is still raging, taking people's lives throughout the world, causing huge economic, financial, and social problems of a large and, at the moment, hardly predictable scale, duration, and harmful consequences. According to the Allianz Risk Barometer 2021 (Allianz Global Corporate & Specialty—AGCS, p. 2),<sup>1</sup> in 2020, the disease caused by coronavirus is among the greatest global business risks and, at the same time, it adversely affects the other business risks in the world, the main ones being: (1) business interruption, (2) COVID-19 pandemic, (3) IT risks related to cybercrime,<sup>2</sup> (4) market developments, particularly the risk of rising insolvency, (5) changes in legislation and regulation, (6) natural catastrophes, (7) fires and explosions, (8) macroeconomic developments, (9) climate changes, (10) political instability and violence, etc. According to Allianz Risk Barometer 2021 (p. 35), the main business risks for small businesses include: coronavirus pandemic, cybercrime, business interruptions, changes in legislation and regulation, market changes, macroeconomic developments, natural catastrophes, fires and explosions, political risks-instability, wars, terrorism, and climate changes. Economy and companies at the global and local levels are all exposed to the listed risks and the adverse impacts (except several sectors such as the pharmaceutical industry, which is less affected by the current crisis; on the contrary, it opens new markets and enlarges the existing ones, increases income and profits). Both in the world and locally, the particularly threatened entities include small and medium-sized enterprises, primarily in the underdeveloped, instable economies such as in Bosnia and Herzegovina which are, due to frequent business interruptions, falling income, disruptions in supply and distribution, lower export of goods/services, unfavorable conditions for obtaining loans, illiquidity, business losses and financial instability, without the adequate help of the state and Entities, exposed to a higher risk of

---

<sup>1</sup> Allianz Global Corporate & Specialty (AGCS) has been conducting research of global business risks over the past ten years based on the analyses and opinions of 2,769 experts (CEOs, experts for risk management, brokers and insurance experts) from 92 countries and territories.

<sup>2</sup> The accelerating digitalization and frequent online operations (due to the pandemic) increase the exposure to IT risks. According to data by US FBI, in April 2020, cybercrime increased by 300%, and cybercrime is now estimated to cost the global economy over a billion of US\$, which is a 50% increase compared to 2018/2019.

closures and bankruptcy. Many micro enterprises in BiH have already disappeared from the market. According to AGCS Barometer for 2021, the five leading risks threatening the financial sector in 2021 are: cybercrime, coronavirus pandemic, business interruptions, changes in legislation and regulation, and macroeconomic changes.

For BiH economy, risk exposure is particularly prominent in terms of the increased market risk (greater uncertainty due to the pandemic and instability, particularly the political instability, deters investors, decreases money inflows from BiH diaspora, etc.), risk of illiquidity of economy, state, and privately owned enterprises, layoffs and increasing unemployment, and the resulting impossibility and/or more difficult repayment/collection of the due insurance premiums, credit, fiscal, and other receivables/debts, which results in illiquidity in the chain and, among other things, increases the (re)insurance company's liquidity risk, delinquency, and the number of non-performing loans, which in turn increases the credit risk, which will inevitably result in the increase of the risk of decreased turnover, income and profits, and losses in the local financial sector,<sup>3</sup> which is already happening in the developed financial markets where the coronavirus pandemic has caused huge losses at stock exchanges and a significant decrease of capital value, as well as an increase of non-performing loans (as a rule, these are loans which are over 90 days overdue). Appendix A in the Appendix of the paper presents an overview of non-performing loans (NPL) in the European banking sector in 2019 and expectations—predictions of a significant increase of NPL ratio in all countries, particularly in Italy (from 6.7 to 20.3) and Spain (from 3.2 to 10.6) due to COVID-19 pandemic.<sup>4</sup> According to the World Bank data, the average share of NPLs in total loans in 100 world countries was 6.01% in 2019. The highest share of NPLs was recorded in Equatorial Guinea, 48.81%, and the lowest in Macao, 0.24%. Within the ranked 100 countries, BiH ranked 22, with recorded 7.41% share of NPLs in the total bank loans. In June 2020, NPL ratio in BiH amounted to 6.7%, and in the first quarter, it amounted to 6.6%. Data in Appendix B in the Appendix reveal that BiH had the highest NPL ratio in December 2020 (21.2%) and the lowest in June 2008 (3.0%).

According to the Insurance Supervision Agency of the Federation BiH, the impact of COVID-19 on the insurance market in this Entity resulted in the 1.6% premium decrease cumulatively, in the period 01.01.2020 to 30.11.2020 compared to the same period of 2019, while the monthly premium November 2020/November 2019 decreased by 2.1%. Only some types of insurance, which have a very low share in the total premium did not experience a decrease in premium, while the main types of insurance experienced a considerable decrease in premium (01 accident insurance experienced a 27.3% decrease, 02 Health insurance experienced a 7.6% premium decrease, 03 Motor hull insurance decreased by 7.1%, 08 Fire and other natural perils insurance had a 26.3% fall, 09 Other damage to property insurance—a fall

<sup>3</sup> See WORLD BANK WESTERN BALKANS GROUP, REGULAR FINANCIAL REPORT no. 17, Spring 2020, Economic and Social Impact of COVID-19.

<sup>4</sup> Source <https://europhoenix.com/blog/may-covid-19-trigger-a-european-banking-crisis-by-les-nemethy-and-nicolas-beguin/NPL> Markets, Forecasting NPL Ratios after Covid-19, May 6, 2020.

of 8.1%), while 07 Goods in transit insurance experienced a growth in premium of 10.6% and 10 Motor third-party liability increased by 8.1%. Type 19, Life insurance, experienced a 12.3% growth compared to November 2019.<sup>5</sup>

The above presented data on the global/local socioeconomic crisis and risks confirm that the coronavirus pandemic caused the global systematic risk, whose duration, scale, and harmful consequences are uncertain.

How to assess risks and damages, how to decrease the level of uncertainty and future damages (caused by this pandemic) in the investment portfolio of the insurance industry in BiH is the topic of this paper.

## 2 Methods and Content of the Research

The paper highlights the application of contemporary methods in managing the (re)insurance company's investment portfolio, primarily the *method for the securities portfolio risk management and the method for credit risk management*.

The paper starts from the valid local regulations pertaining to insurance—the adjusted Solvency I regime, as well as the prescribed methodology for main risks assessment and for the quantitative analysis of the investment portfolio's impact on BiH (re)insurance companies' solvency.

The assessment of the *securities portfolio risk* used the model of variance and standard deviation, which reveal the degree of deviation of the security yield from the expected mean value of return, while the yield (positive or negative, or equaling zero) is measured by the increase, decrease, or keeping the price of the security at the same level between two measurement periods. The correlation of yield between two securities is measured by the correlation coefficient and covariance. The usual *models for assessing a potential investment loss resulting from the systematic risk* include: capital asset pricing model (CAPM) and models for assessing value at risk (VaR): historical method, Monte Carlo method, and parameter method.

## 3 Business Goals and Results—Comparative Data on the Development of the Insurance Sector in BiH and Surrounding Countries

(Re)insurance companies primarily achieve the set economic-business goals from their basic activity, i.e. negotiating deals—selling life and nonlife insurance policies. Besides, insurers generate a significant part of income from activities of investing

---

<sup>5</sup> Source COVID-19 i tržište osiguranje u FBiH na 30.11.2020. <http://www.bosnare.ba/vijest/covid-19-i-trziste-osiguranja-u-fbih-na-30-11-2020-103>.

their own and temporarily free “someone else’s” resources (technical reserves<sup>6</sup>), pursuant to law. The purpose of technical reserves is solely settling obligations from the signed insurance contracts. The valid regulations in BiH and its Entities strictly regulate the purposes and levels at which resources for covering technical reserves or assets for covering the mathematical reserve of the insurance company can be invested and deposited. Like other investors, insurance companies generate their investment portfolio by business transactions, buying and selling different financial instruments in financial markets.

In their business, every company, including every (re)insurance company, is as capable of surviving and being successful in the market as it is capable of ensuring the long-term *solvency* and current and long-term *liquidity*, to be able to service its obligations due in the negotiated (prescribed) terms and, on top of it, to run a *profitable business* and thus preserve and enlarge its initial capital. Therefore: long-term fulfillment of prescribed conditions for solvency (having at disposal the optimum reserves in the amount sufficient to cover its liabilities, the minimum prescribed guarantee fund, ensuring and maintaining the solvency margin as the minimum amount of capital for covering the unexpectedly high liabilities for claims); steady insurance sales, i.e. market share; timely collection of receivables in line with their maturity dates; stable sources of financing; stable, certain, and liquid placements of temporarily free resources of the insurance company, are essential prerequisites for its successful business and development. Thanks to the possibility to realize the investment insurance funds in a short time, without a significant decrease in their value, insurance companies that manage to achieve all this are stable and liquid, while the generated financial revenues increase profits from the insurance business and contribute to the expansion of service range and faster settling of claims and, indirectly, to the increase of the number of insurance clients and revenues from premiums, as well as to an increase of the company’s market share (circular inter-dependent flow; solvency, liquidity, profitability, market-profit potential).<sup>7</sup> The fundamental priorities or the sequence of goals-principles of (re)insurance company’s investment are (1) liquidity—liquidating securities within a short time and at a fair value, (2) security and satisfactory dispersion of funds placement pursuant to regulations, (3) satisfactory yield on the invested funds, as a rule at the level of the weighted cost of the company’s own and borrowed capital (WACC model), which includes the rate of yield on the investment without risk incremented by the so-called risk premium (yield above the non-risk rate), i.e. at least at the level of the average interest rate in the capital market, (4) ensuring the stability of the company’s assets and, on the other hand, reliance on the stable—long-term sources of financing the company. These principles and priorities derive from the (re)insurer’s need to prevent the investment of the temporarily free insurance funds from threatening its own current and long-term liquidity and solvency,

---

<sup>6</sup> Technical reserves of the insurance company include: mathematical reserve of life insurance, transferable premium (instead of mathematical reserve) for nonlife insurances, reserves for reported claims, reserves for unreported claims, reserves for bonuses and rebates, claims equalization reserves, and other reserves related to the insurance business.

<sup>7</sup> For more details, see the book by Ž. Šain, E. Taso: Due diligence – Procjena vrijednosti osiguravajućeg društva, EFS UNSA, Sarajevo, 2015. p. 114.

as well as the fulfillment of due liabilities. For this reason, *(re)insurance companies, as a rule, tend to invest their funds in more liquid and certain placements that carry lower returns-yields*, than in riskier investments with a higher rate of return. Together with the optimum structure of insurance portfolio harmonized with market needs, the key prerequisite for efficient investing is the developed financial market with organized supply and demand of cash and diverse financial instruments, primarily state securities.

In the developed market economies, possibilities for investment are far greater and more diverse due to the more developed financial market with a supply of a large number of different kinds of financial instruments. At the same time, insurance companies in developed economies show considerably more attention and corporate responsibility in maintaining the liquidity and profitability of investing the resources of technical reserves and guarantee fund in line with the new regime Solvency II (in the EU). The insurance sector is the largest institutional investor in Europe (with assets worth over 10 trillion euros), which is highly concentrated in a small number of countries (the United Kingdom, France, Germany, Italy). EIOPA's data (Q2 2019) show that the traditional portfolios of the European insurance industry consist of government bonds (31.4%), corporate bonds (32.2%), listed and unlisted equity (15.1%), cash and deposits (5.2%), mortgages and loans (5.7%), property (2.2%), and other assets (8.2%). *Source* [https://www.ecmi.eu/sites/default/files/tfaa\\_final\\_report\\_ecmi.pdf](https://www.ecmi.eu/sites/default/files/tfaa_final_report_ecmi.pdf). Appendix C in the Appendix shows the growth of the investment portfolio of the insurance in 32 European countries in the period 2010–2018, by 32.34% or, in billion euros, from 7,697.39 in 2010 to 10,186.31 in 2018, while Appendix D in the Appendix shows the growth by basic insurance types.

The insurance sector in BiH experienced a continuous growth of assets in the period 2015–2018 and ranks second (after banks) by its share in the total assets of BiH financial sector. However, *the total potentials of local financial institutions are very limited and modest compared to the region and Europe*, which is shown in Appendix E in the Appendix. Besides, investments of BiH insurance companies are many times smaller by value and differ significantly by structure from the investments in the developed European economies, which is presented in Appendix F for Entity of the Federation BiH.

The low competitiveness of the insurance sector results from the overall unsatisfactory macroeconomic position of BiH. Compared to the developed European countries, BiH economy and the insurance sector within it are insufficiently developed and considerably lag behind the average of the European Union member countries by: gross domestic product, generated premiums, the share of total premium in the gross domestic product, GDP per capita, and the premium amount per capita (Tables 1, 2, 3, and 4). Indicators for insurance penetration and insurance density, i.e. the average annual gross premium per capita also reveal the insufficient development of the BiH insurance market compared to the average for the European Union, which can be seen from the following data:

By the value of assets, capital, and generated insurance premium, the insurance sector in BiH lags behind the insurance sector in the countries of the region, and particularly behind the more developed insurance industry in the EU member countries.

**Table 1** Trend of GDP and premium in EU27 and BiH in the period 2007–2019

	2007	2008	2010	2015	2018	2019
GDP in EU 27 (in million euros)	12,398,526	12,494,352	12,280,644	14,710,626	18,118,140	15,376,490
<i>Index previous year = 100</i>		100.77	98.29	119.79	123.16	84.87
GDP in BiH (in million euros)	11,098	12,675	12,692	14,435	17,132	18,066
<i>Index previous year = 100</i>		114.21	100.13	113.73	118.68	105.45
GDP BiH/GDP EU 27	0.09%	0.10%	0.10%	0.10%	0.09%	0.12%
Premium in EU 27 ( million euros)	1,026,645	1,066,731	1,116,225	1,218,895	1,315,377	1,047,139
<i>Index previous year = 100</i>		103.90	104.64	109.20	107.92	79.61
Premium in BiH (million euros)	206	231	242	306	366	390
<i>Index previous year = 100</i>		112.14	104.76	126.45	119.61	106.56
Premium BiH/Premium EU 27	0.02%	0.02%	0.02%	0.03%	0.03%	0.04%

Sources Derived from the data by Eurostat newsrelease euroindicators <http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do> and data by Insurance Agency of BiH Insurance Market Statistics for 2007–2019

**Table 2** Premium per capita and share in GDP in EU27 and in BiH

	2007	2008	2010	2015	2017	2018	2019
Premium per capita EU27 eur	2,204	2,540	2,241	2,390	2,348	2,335	2,122
Premium per capita in BiH eur	54	60	64	80	100	135	112
Share of premium in GDP for EU27 (%)	8.80	8.60	8.43	8.29	7.20	7.26	6.81
Share of premium in GDP for BiH (%)	1.90	1.80	1.91	2.12	2.18	2.11	2.19

Derived from: Insurance Agency of BiH Insurance Market Statistics – reports for 2007, 2008, 2009, 2010, 2015, 2019

Besides, both important financial sectors in BiH (banking and insurance) considerably lag behind the developed EU countries and most countries in the region by financial potential.

The sources of the BiH economic lag, including that of the insurance sector, behind the countries in the region and, far more, compared to the developed EU members

**Table 3** Premium per capita in EU27, in BiH, and in neighboring countries

	2007	2008	2010	2015	2017	2018	2019
Premium per capita EU 27 eur	2,204	2,540	2,241	2,390	1,465	2,335	2,122
Premium per capita Croatia eur	278	302	292	270	294	326	341
Premium per capita Serbia eur	75	80	80	93	112	121	131
Premium per capita Montenegro eur	82	96	99	124	131	140	152
Premium per capita in BiH eur	54	60	64	80	100	135	112

Derived from the data by the Insurance Agency of BiH Insurance Market Statistics – reports for 2007-2019 [https://europa.eu/european-union/about-eu/countries/member-countries\\_hr](https://europa.eu/european-union/about-eu/countries/member-countries_hr)

**Table 4** Position of BiH economy<sup>8</sup> by some global competitiveness indicators

	2007/08	2008/09	2009/10	2010/11	2013/12	2016/15	2019/2018
	of 131	of 134	of 133	of 139	of 144	of 140	of 141
Global competitiveness index	106	107	109	102	88	111	92
Institutions	113	123	128	126	85	127	114
Macroeconomic stability	90	57	69	81	97	98	64
Market size	80	92	90	93	93	97	101
Efficiency of commodity markets	113	123	125	127	109	129	108
Technological readiness	110	109	95	85	68	79	80
Innovation and sophistication	123	129	127	120	99	120	117

<sup>8</sup> Izvještaj o kompetitivnosti BiH 2019–2018, Prof. dr. Zlatko Lagumdžija, Ekonomski fakultet Sarajevo, Sarajevo, December 2016.

and their average, *regardless of the coronavirus pandemic*, result from the insufficient post-war investments in economy, constant political instability, and provoking crises, from closures and disappearance of numerous business entities and their numerous subcontractors in the war and the post-war period, high unemployment, low wages, increase in prices of goods and services, decrease in the population's standard and purchasing power. From once successful, export-oriented economy of BiH that used to achieve a continuous growth and employ over a million workforce, after the 1992–1995 war until now only a disoriented and disintegrated economic system remained, very vulnerable to a smallest crisis, with the accumulated, insatiable administration, and social benefits above capacities.



## 4 Management of Insurance Company’s Securities Portfolio

Each security (VP) has two main characteristics:

- *Yield on security (VP)* that the investor receives from investing funds in a certain VP (*bonds* and *granted loans* generate the yield in the form and amount of the negotiated *interest*, while in the case of *shares*  $Yield\ on\ a\ share = Dividend + Capital\ gain$ )
- *Risk to VP*—accompanies each investment under uncertainty (*the measure of risk is variance ( $\sigma^2$ ), which represents the weighted square deviation of random yields of a given VP from the expected (average) yield rate of this VP; the alternative measure is standard deviation ( $\sigma$ ), which is calculated as the square root from variance:  $\sigma = \sqrt{\sigma^2}$* ).

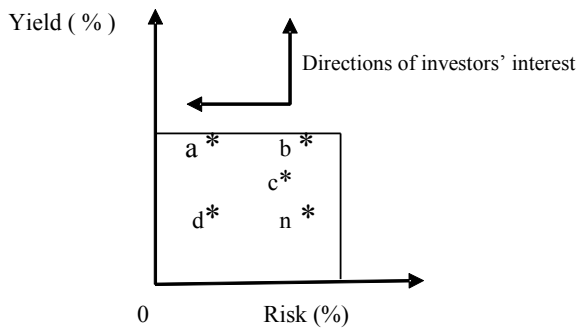
Besides yield and risk, the selection of the satisfactory portfolio depends on two more important factors:

- *Correlation coefficient (statistical correlation of yields) between individual securities in the portfolio, and*
- *Investor’s benefit preferences, i.e. the degree of their risk appetite or aversion (Some investors prefer higher yields and are willing to accept even an increased risk. Others are more risk averse than the average investors, and therefore typically invest in low-risk securities or in government bonds).*

Using the combination of these criteria, it is possible to form an *efficient securities portfolio* which, at a certain level of risk, brings the highest yield rate, i.e. a portfolio which carries the lowest risk for the expected yield rate.

The aim of investing the temporarily free monetary resources in different financial instruments is *gaining profit*. By doing so, the investor takes a certain *investment risk* (Graph 1).

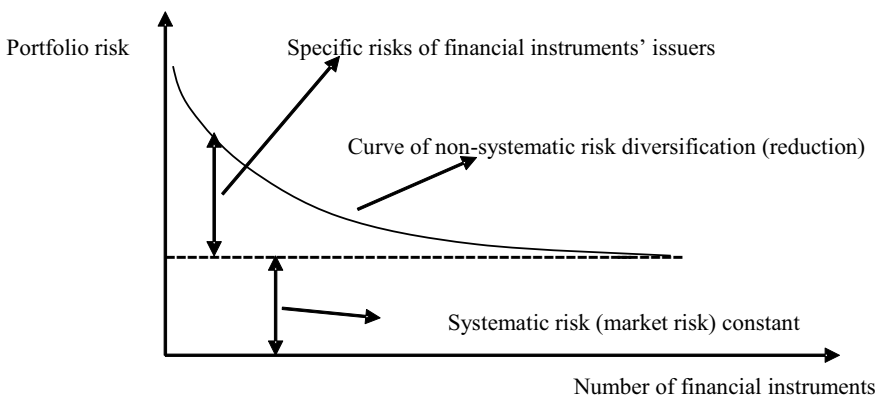
**Graph 1** Interdependence of the expected return and loss risk in securities



All types of risk are characterized by the uncertainty of the occurrence of an unfavorable event in the future, which can cause financial losses for investors. The crucial risks that (re) insurers as investors are exposed to in financial markets are: *insurance risk* (which includes risks based on negotiating-selling policies for 18 nonlife insurances and life insurance), *liquidity risk* (impossibility to liquidate securities and other investments in a short term at a fair value to pay its liabilities due), *concentration risk* (which includes all risk exposures where the potential loss is so high that it threatens the (re)insurance company’s solvency or financial position), *market risk* (risk or loss or an unfavorable change in financial conditions due to changes in the credit position of financial instruments’ issuer, other contractual parties, and any debtors that the (re) insurance companies are exposed to in the form of the concentration of the risk of the other contractual party’s failing to fulfill its obligations, yield risk, or market risk), *operational risk* (risk of a loss due to inadequate internal business processes, employees, system, or external events) (Graph 2).

The selection of the optimum securities portfolio is guided by the rules of *Markowitz portfolio theory*, which includes the analysis of securities, portfolio analysis and selection, and periodical measurement of the selected portfolios’ results. Its starting assumptions are as follows:

- Investors make decisions under uncertainty and they are interested in both a higher yield and decreased risk of the securities portfolio;
- Investors prefer higher compared to lower yield;
- The basic way to increase yield is to take greater risk;
- The efficient securities portfolio is one that carries the highest expected yield (return) for the same level of risk, i.e. a portfolio that carries the smallest risk for the same level of the expected return;
- The key activity for decreasing a non-systematic risk is diversification of investment into several different financial instruments.



**Graph 2** The structure of the total portfolio risk consists of systematic and non-systematic risks

The structure of the total risk of the (re) insurance companies' (and other investors') securities portfolio consists of the *specific or non-systematic risks* to individual financial instruments, i.e. risks deriving from their issuers' business performances (such as achieved sales, company's revenues, profits, liquidity and financial stability, its competitiveness, market share, capacity utilization, quality of management and staff, kind of business, technology, sales range, etc., which affect prices of the issued securities, i.e. yields and risks, their saleability, etc.), and the *systematic risk or market risk*. The graph above reveals that the curve of the non-systematic risk to securities portfolio decreases with the diversification of investment in a large number of different financial instruments. The full effect of diversification is particularly achieved if the securities that are invested in *have the mutually negative correlation coefficient* (move in opposite directions), so that the loss on one security is accompanied with the gain on another, which decreases the volatility<sup>9</sup> of the whole portfolio. On the other hand, if securities are mutually positively correlated (if they move together in the same direction), the risk to such a portfolio increases.

*Diversification of investment* in several different financial instruments, i.e. the optimization of their share in the portfolio according to the criteria of yield, risk, and mutual securities correlation in the portfolio, *can decrease and even completely avoid non-systematic securities risks*, i.e. risks related to their issuers' business and results, although it *does not avoid or reduce the market or systematic risk*. Due to the inevitable impact of the ever-present systematic risk, *the total risk to the portfolio* of the (re)insurance company composed of several financial instruments cannot be fully neutralized. Studies in the developed market economies established that effects of diversification, i.e. of the total decrease of specific-non-systematic portfolio risk, is already achieved with investment of resources in 10 to 20 financial instruments.<sup>10</sup> With respect to the relationship between risk and the number of shares in the portfolio, authors Elton and Gruber believe that 51% of the standard portfolio deviation is eliminated with an increase in the number of shares from 1 to 10; adding 10 more shares further eliminates 5% of the standard portfolio deviation, while an increase of the number of shares to 30 additionally eliminates only 2% of the standard deviation–portfolio risk.<sup>11</sup>

Systematic risk is a result of the changes in general socioeconomic and political conditions due to the impact of different sources (at present, it is due to the current COVID-19 pandemic) that affect all players and the whole capital market, particularly the conditions in supply and demand, changes of securities' prices and yields,

<sup>9</sup> Volatility is a measure of an uncertain change of a variable in a period of time. It shows the size of change in the price, i.e. the standard deviation of the change in security price over a given past period. The greater the volatility, the greater its risk.

<sup>10</sup> Source Evans, J. L., Archer, S. H. (1968): Diversification and the Reduction of Dispersion: An Empirical Analysis, *Journal of Finance*, Vol. 23, No. 5, pp. 761–767.

<sup>11</sup> Source Elton, E. J., Gruber, M. J. (1977): Risk Reduction and Portfolio Size: An Analytical Solution, *Journal of Business*, No. 50, pp. 415–437.

interest rates, currency values, inflation, political stability, business interruptions, bankruptcies, unemployment, etc. Market risk in a broader sense includes: risk of a change in the securities’ price, interest rate risk, currency risk, and the risk of a change in commodities’ price. Some authors believe that the concept of market risk primarily implies the risk of change in securities’ prices in financial markets, while they classify and analyze other market risks separately.

*The risk of the value at risk from investing due to the unfavorable impact of systematic risk* is typically assessed using the so-called capital asset pricing model (CAPM), and models for assessing value at risk (VaR)—historical method, Monte Carlo method, and parameter method.

*Achieving the desired level of balance in every investor’s investment portfolio also depends on: their utility preference, their willingness/ability to take a certain level of investment risk with respect to the expected yields, the applied business strategy for achieving economic goals and limiting the investment risk exposure, on the availability, i.e. supply of different financial instruments in the financial market, general market conditions, interest rate levels, the level of national economy development, and on economic conditions and trends at the regional and global level (stability, economic growth and prosperity, expansion or recession, increase–decrease of gross domestic product, deceleration-acceleration of economic activities, decline-growth of employment and consumption, etc.).* At the international level, specialized agencies such as *Moody* and *Standard & Poors* determine and assess the scales of the risk of investing in different securities.

Risk scales-securities rating by specialized agencies

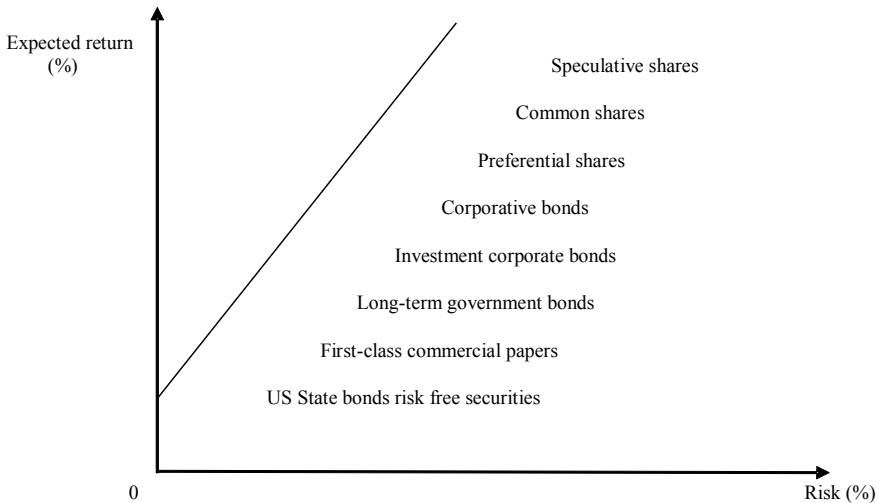
MOODY	STANDARD & POOR’S
Aaa Highest quality	AAA Highest rating
Aa High quality	AA High rating
A Upper-medium grade	A Upper-medium rating
Baa Medium grade	BBB Medium rating
Ba Have speculative elements	BB Speculative rating
B Does not have characteristics of target investment	B Very speculative rating
Caa Poor quality: likelihood of being near or in default	CCC-CC Extremely speculative
Ca Highly speculative: often in default	C Reserved for income from bonds with nonpaid interest
C Lowest quality	DDD-D In default

The lines marked in red—risk levels are the highest quality securities, while other categories include securities with an increased, great, and high risk (Graph 3).

The risk of loss from investment increases proportionately with the increase of the expected yield from the investment, which is illustrated by the following graph.

### 5 Managing the Insurance Company’s Credit Portfolio

The risk of a loan/credit that the insurance company (and/or another lender) granted to a third legal or natural entity for business purposes (financing permanent and/or working assets) or for consumption, under the commercial repayment terms includes the risk of default in loan obligations (repayment of the due principal and/or negotiated interest) by debtor-borrower in the term and amounts according to the negotiated credit agreement and payment schedule, due to which the insurance company as a lender can incur a loss and additional costs. In a broader sense, besides the basic risk of losing the principal and interest, and costs of forced collection of the loan, the credit risk also includes the risk of losses based on the lost new turnover of delinquent principal and interest due to the debtor’s default in repayment, then the risk of potential lender’s illiquidity, and impossibility to service due obligations based on claims and other bases and to creditors.



**Graph 3** Expected yield rate and risk of different securities<sup>12</sup>

<sup>12</sup> Prof. dr Dragana Đurić, Uvod u finansijski menadžment, p.19.

The non-performing loans (NPLs) cause a series of problems in the chain: they adversely affect insurance companies as lenders, i.e. commercial banks and other creditors causing problems in their liquidity in servicing their own obligations, decrease their profits, generate high losses based on write-offs and/or sales of NPLs as irrecoverable assets, thus causing the increase of reserves for covering potential losses based on non-performing loans, decrease of available resources for new lending based on the unrepaid principal (lost new turnover). All this results in the deceleration of the growth and development of the entire economy.

In this respect, the prevention of non-performing loans granted to legal and natural entities, particularly lending for economic investments deserves a particular attention, i.e. the application of a complete methodology that encompasses the analysis of business and results of the loan applicant over the past three to five years, analysis of the new investment for which the loan is requested, supply and demand in the market, development plans and possibilities to repay the debt, needs and resources of secondary security, and providing a collateral in the course of repayment of loan obligations.

Before approving the loan to the applicant, the creditor has to check the loan application and appendices and analyze factors that the company-loan applicant's creditworthiness depends on. As a rule, credit analysis comprises *five elements (the so-called 5C): applicant's character, their capital, production and financial capacity, collateral for the loan, and conditions of the applicant's business*. Since loan obligations can be repaid from profits, income or sales of part of assets or securities, and from other loans, or from additional capital, the lender is primarily interested in the borrower's business liquidity, stability and profitability, turnovers of customers, supplies and suppliers, the degree of indebtedness, the degree of the loan debt coverage, cash flow, balance sheet and income statement when considering punctual repayment of the loan (together with monitoring and regular reporting on the implementation of the business plans for production and sales, market share, employment). Besides the mandatory *assessment of the vulnerability of the investment project and loan beneficiary-debtor* (e.g. to cases of an unpredicted decline in sales, problems with collection, providing raw materials or energy products under the planned terms of acquisition and payment, etc.), it is mandatory to analyze the economic justification of lending to the specific economic investment, and particularly: *internal profitability rate, current net value of the project, period of return on investment, sensitivity-profitability threshold or the break-even sales level of the investment, as well as the set of mandatory appendices to the loan application, and other issues*.

Besides the assessment of the investment plan, of the investor-loan beneficiary's past, current, and future overall business, the creditor should insist on the simultaneous (if possible, *majority*) *participation in the investment and the debtor-loan beneficiary itself*. Together with the covered credit risk with the appropriate level of interest proportionate to the taken credit risk, it is necessary to ensure/negotiate and register, at court or with other authorized state authorities, *means for ensuring punctual repayment of loan obligations* (suitable mortgage, guarantee, assurance, etc.).

The significance of ensuring additional guarantees for the punctual servicing of loan obligations in BiH is great, particularly due to the fact that the valid legislation in BiH and Entity regulations regulate that, for the case of business interruptions that result in the discontinued operations due to liquidation or bankruptcy, the company is accountable for its remaining liabilities with its assets up to the amount of its founding capital. In the limited liability companies in BiH, it amounts to a couple of thousand KM while, at the same time, such a debtor's liabilities sometimes exceed multi-million amounts. Besides, assets and goods of its founder-owner and entities affiliated with him are exempted from the obligations of such a company-debtor.

*Regular monitoring of the repayment of loan obligations and of the debtor's business* is important because of the rule that a non-performing loan is preceded by the so-called warning signs—potential problem indicators (financial: in bookkeeping and periodical financial statements of the loan beneficiary-debtor, such as illiquidity, losses, over-indebtedness, slow turnovers of receivables and supplies uncoordinated with liabilities maturity, large cash withdrawals, high fixed costs, particularly interest on additional borrowing; managerial: unfilled crucial staff, frequent changes of managers, CEO and owner are one person, ethical characteristics—wastefulness and similar occurrences; operational: hampered provision of quality raw materials, goods, energy products, impaired relations with suppliers, decline in revenues, increase of variable costs due to rejects, breakage and malfunctions, low-quality raw materials, defective equipment, untrained labor, sales of permanent assets, change in activity, impossibility to obtain financial statements, tardy bookkeeping, etc.; banking: frozen accounts, increasing indebtedness, newly taken unpredicted loans, etc.). Due to the above listed, the basis of preventing a granted loan from non-performance is establishing, in the loan agreement, the right and obligation of periodical reporting on spending loan funds for agreed purposes and according to the agreed dynamics, creditor's right to directly visit the debtor and control its business and results, including the credit beneficiary's periodical obligation to submit reports on its business, on the condition of collateral and other negotiated means for ensuring punctual repayment of the principal and loan interest.

## 6 Conclusion

In the conditions of the current crisis, it is particularly important to apply the described rules of investing into securities and of lending. Besides liquidity, certainty, profitability, and diversification of the securities portfolio, (re)insurance companies should ensure their own liquidity in this time of crisis primarily by the *preference for granting short-term loans* mostly to stable and profitable clients, which finance the majority of the investment by their own means and provide the adequate collateral/guarantee for the loan.

## Appendices

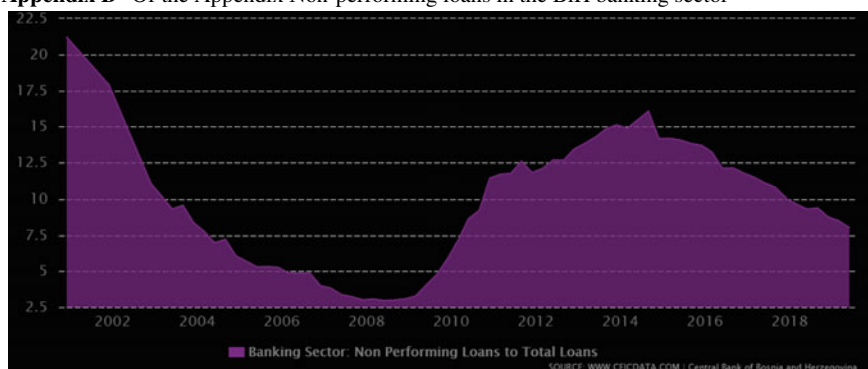
See Appendices A, B, C, D, E and F.

### Appendix A of the Appendix Non-performing loans in the European banking sector

Country	Historical Max NPL Ration (%)	Latest NPL Ratio (%)	Covid Max NPL Ratio (%)	Percentage increase	Total loans 2019 bn EUR	Covid Max NPL bn EUR
Austria	3.5	1.6	7.1	333	498.0	35.0
Czech Republic	29.3	3.1	8.0	155	138.0	11.0
France	6.3	2.5	7.5	199	4,819.0	360.0
Germany	5.2	1.3	3.3	155	2,394.0	79.0
Hungary	16.8	1.5	6.0	295	60.0	4.0
Italy	18.1	6.7	20.3	202	1,731.0	351.0
Poland	21.22	3.9	8.1	110	133.0	11.0
Slovak Republic	31.6	2.9	6.0	119	43.0	3.0
Spain	9.4	3.2	10.6	234	2,450.0	259.0
United Kingdom	4.0	1.3	4.1	218	5,544.0	229.0

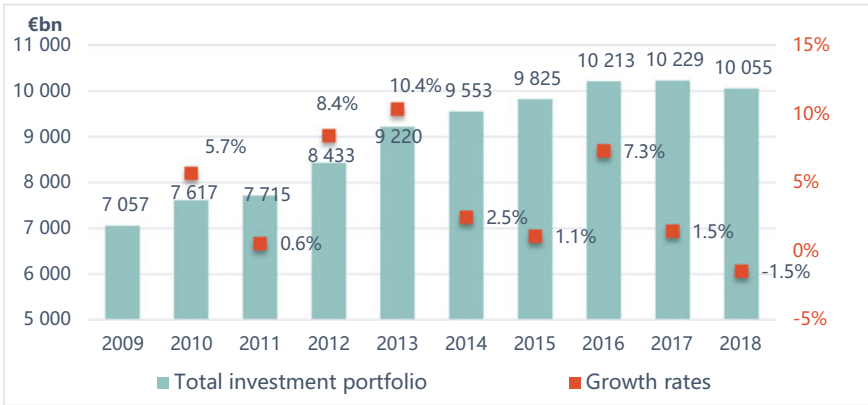
According to the trends in the latest NPL ratio available (ratio of 2019), COVID-19 will have a great impact on NPL, particularly in Italy and Spain, where NPL ratio could exceed 20 and 10%, respectively. *Source* <https://europhoenix.com/blog/may-covid-19-trigger-a-european-banking-crisis-by-les-nemethy-and-nicolas-beguin/NPL> Markets, Forecasting NPL Ratios after Covid-19, May 6, 2020

### Appendix B Of the Appendix Non-performing loans in the BiH banking sector





**Appendix C** Of the Appendix Total investment portfolio of the insurance sector in 32 European countries 2009-2018 (€bn)



Source European Insurance in Figures - 2018 data, Insurance Europe, 2020, p.42

**Appendix D** Investment portfolio of life and nonlife insurance in 32 European countries

	2009	2010	2011	2015	2016	2017	2018
<b>Non-life insurers' investment portfolio</b>							
Sample	1 198	1 251	1 283	1 593	1 661	1 658	1 617
% Change		2.1%	2.2%	2.5%	5.9%	1.0%	-2.4%
<b>Life insurers' investment portfolio</b>							
Sample	5 320	5 796	5 842	7 465	7 685	7 785	7 647
% Change		6.5%	-0.1%	0.4%	7.0%	2.8%	-1.5%

Source European Insurance in Figures - 2018 data, Insurance Europe, 2020, p.42

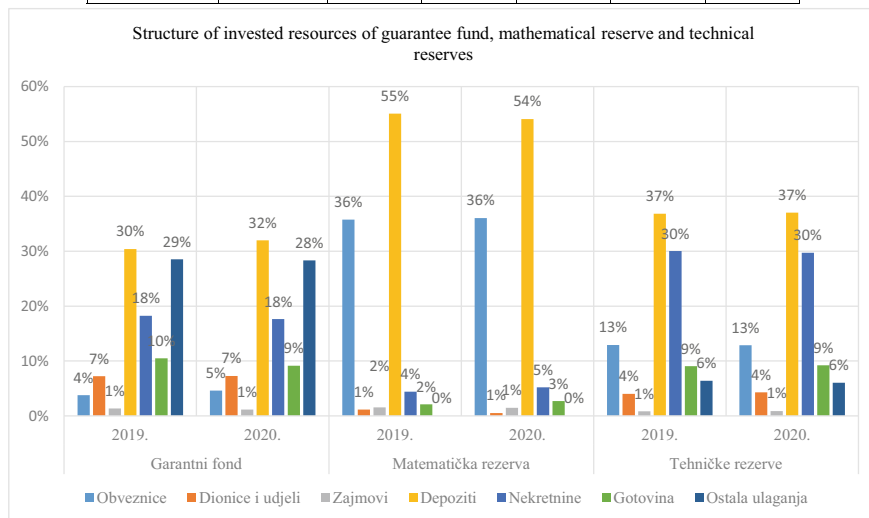
**Appendix E** Structure of financial institutions' share in the BiH financial market

million KM	2015		2017		2018		2019		Growth index	
	Assets	Share (%)	Assets	Share (%)	Assets	Share (%)	Assets	Share (%)	18/17	19/18
Financial institutions	23,829	87.47	27,249	88.26	29,854	88.46	32,508	88.70	109.6	108.9
Banks	834	3.06	855	2.77	889	2.63	855	2.33	104.0	96.2
Investment funds	1,466	5.38	1,717	5.56	1,819	5.39	1,967	5.37	105.9	108.1
Insurance and reinsurance companies	640	2.35	791	2.56	891	2.64	996	2.72	112.6	111.8
Microcredit organizations	475	1.74	260	0.84	297	0.88	324	0.88	114.2	109.1
Leasing companies	27,244	100	30,872	100	33,750	100	36,650	100	109.3	108.6
Total for sector										

*Data source* Statistics of insurance market in BiH for 2019 (July 2020).

### Appendix F Value and structure of investments by insurance companies in the Federation BiH in 2019 and 2020

	Guarantee fund		Mathematical reserve		Technical reserve	
	2019	2020	2019	2020	2019	2020
Total investment	234,558,219	251,262,812	555,867,912	574,365,926	404,642,437	414,118,873



Source Report on the insurance sector of the Federation BiH by the Insurance Supervision Agency of F BiH

## References

- Allianz risk barometer (2021) (Allianz Global Corporate & Specialty-AGCS)
- Elton EJ, Gruber MJ (1977) Risk reduction and portfolio size: an analytical solution. *J Bus* (50)
- Evans JL, Archer SH (1968) Diversification and the Reduction of dispersion: an empirical analysis. *J Financ* 23(5)
- BiH Competitiveness report 2019–2018
- NPL Markets (2020) Forecasting NPL Ratios after Covid-19, May 6
- Đurić D (2006) *Uvod u finansijski menadžment*, Europrint, Beograd
- WORLD BANK GROUP (2020) Western Balkan Report no. 17. Spring. Economic and social impact of COVID-19
- Šain Ž, Taso E (2015) *Due diligence–Procjena vrijednosti osiguravajućeg društva*, EFSA, Sarajevo