Springer Proceedings in Mathematics & Statistics

Alberto A. Pinto
David Zilberman *Editors*

Modeling, Dynamics, Optimization and Bioeconomics II

DGS III, Porto, Portugal, February 2014, and Bioeconomy VII, Berkeley, USA, March 2014 - Selected Contributions

Springer

Springer Proceedings in Mathematics & Statistics

Volume 195

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Alberto A. Pinto · David Zilberman
Editors

Modeling, Dynamics, Optimization and Bioeconomics II

DGS III, Porto, Portugal, February 2014, and Bioeconomy VII, Berkeley, USA, March 2014 - Selected Contributions

Springer

Editors
Alberto A. Pinto
Department of Mathematics
and LIAAD-INESC TEC
University of Porto
Porto
Portugal

David Zilberman
Department of Agricultural
and Resource Economics
University of California
Berkeley, CA
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-55235-4 ISBN 978-3-319-55236-1 (eBook)
DOI 10.1007/978-3-319-55236-1

Library of Congress Control Number: 2017943186

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, express 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.

Printed on acid-free paper

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

Alberto Pinto dedicates this volume
to Maria Barreira Pinto.

David Zilberman dedicates this volume
to Leorah Abouav-Zilberman.

Foreword

The aim of this book "Modeling, Dynamics, Optimization and Bioeconomics II" follows the aim of the book "Modeling, Dynamics, Optimization and Bioeconomics I," that is the exploration of emerging and current cutting-edge theories and methods of modeling, optimization, dynamics, and bioeconomy. The theories and techniques presented here originated from dynamics, statistics, control theory, computer science, and informatics and are applied to novel and innovative real-world applications. During the past decades, the use of dynamic systems, control theory, computing, data mining, machine learning, and simulation has gained the attention of numerous researchers from all over the world.

Admirable scientific projects using both model-free and model-based methods coevolved today at research centers and are introduced in conferences around the world, yielding new scientific advances and contributing to the solution of important real-world problems. One important area of progress is the bioeconomy, where advances in life sciences are used to produce new products in a sustainable and clean manner. In this book, scientists from all over the world share their latest insights and important results in the field.

We are very thankful to the editors, Alberto Adrego Pinto and David Zilberman, for having given these experts the opportunity and honor of publishing their contributions. We express our gratitude to them for having prepared a premium work of a remarkable scientific and social value.

Ankara, Turkey
April 2016

Gerhard-Wilhelm Weber

Acknowledgements

We thank the authors of the chapters for having shared their vision with us in this book, and we thank the anonymous referees.

We are grateful to Gerhard-Wilhelm Weber for contributing the foreword of this book.

We thank the executive editor for mathematics, computational science and engineering at Springer-Verlag, Martin Peters, for his invaluable suggestions and advice throughout this project. We thank Ruth Allewelt at Springer-Verlag for her assistance throughout this project.

We thank João Paulo Almeida, Susan Jenkins, José Martins, Abdelrahim Mousa, Bruno Oliveira, Diogo Pinheiro, Filipe Martins, Renato Soeiro, Ricard Trinchet Arnejo, and Yusuf Aliyu Ahmad for their invaluable help in assembling this volume and for editorial assistance.

Alberto Adrego Pinto would like to thank LIAAD-INESC TEC and gratefully acknowledge the financial support received by the FCT—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology)—within project UID/EEA/50014/2013 and ERDF (European Regional Development Fund) through the COMPETE Program (operational program for competitiveness) and by National Funds, through the FCT within Project "Dynamics, optimization and modelling" with reference PTDC/MAT-NAN/6890/2014, and Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF). Alberto Adrego Pinto also acknowledges the financial support received through the Special Visiting Researcher Program (Bolsa Pesquisador Visitante Especial—PVE) "Dynamics, Games and Applications," with reference 401068/2014-5 (call: MEC/MCTI/CAPES/CNPQ/FAPS) at IMPA, Rio de Janeiro, Brazil.

Contents

Economics of Natural Resource Utilization - the Case of Macroalgae

Ruslana Rachel Palatnik and David Zilberman

Abstract The bioeconomy utilizes living organisms and processes them to produce of food, fuel, fine chemicals, and other substances. Macroalgae (seaweed) are promising feedstocks for energy and chemical products while sequestering carbon. A few species are already used as food products or supplements. There is a need for methodologies for economic and policy analysis of novel bioeconomy technologies, taking into account environmental side effects and physical and economic uncertainties. Farmers growing cellulosic energy crops face significant revenue uncertainty due to both production and price uncertainties. The literature reports a wide range of growth rates for macroalgae and the few business case studies show mixed results in terms of production frontier and profitability considerations. This paper contributes to existing literature on ex-ante assessment of algae-based biofuel production. The study points at both the scientific and economic challenges that require multidisciplinary effort to develop viable technologies, cost-effective harvesting equipment/techniques and processing facilities, and supporting infrastructure, as well as to, create the markets for novel, sustainably produced goods.

Keywords Macroalgae · Economic potential · Profitability · Net Present Value

R.R. Palatnik (✉)
Department of Economics and Management, The Max Stern Yezreel
Valley College, Jezreel, Israel
e-mail: rusalik@gmail.com

R.R. Palatnik
NRERC- Natural Resource and Environmental Research Center,
University of Haifa, Haifa, Israel

D. Zilberman
Agricultural and Resource Economics Department, UC Berkeley,
Berkeley, CA, USA
e-mail: zilber11@berkeley.edu

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_1

1 Introduction

The bioeconomy provides a possible solution for the demand to natural resources by substitution of the unrenewable resources with resources derived from biomass [13]. The bioeconomy consists of complex supply chains that include biomass production, transportation, conversion into products at biorefineries, and distribution. One of the major challenges is developing economic decision-making tools to assess novel biotechnologies that incorporate the complex multi-level systems including environmental implications and uncertainties about feedstock production, refining technologies, markets, and policies.

Alternative biomass supply can come from micro- and macroalgae. Microalgae have been the focus of intense research in the last 50 years. However, cost-effective cultivation, harvesting and dehydration difficulties currently prevent broad scale, sustainable microalgae technologies implementation [22]. Marine macroalgae, also ranked among the most efficient photosynthetic organisms on earth, bear valuable chemical compounds [17]. In a parallel vein, in the recent years, macroalgae have been considered a "third or even fourth generation" biofuel feedstock [34].

The rapidly developing technology for cultivating and refining Macroalgae (mainly red, green and brown algae – seaweed groups) draw the attention of biologists (e.g. [24]) and bioengineers (e.g. [8, 20, 36, 52]). Macroalgae, which contain very little lignin and do not compete with food crops for arable land or potable water, have stimulated renewed interest as additional candidates for future sustainable food, platform chemicals and fuel (biofuel) feedstocks.

However, to date macroalgae still account for only a tiny percent of the global biomass supply with $\sim 17 \times 10^6$ wet weight (WW) ton of macroalgae in comparison to 16×10^{11} tons of terrestrial crops, grasses and forests [42, 43, 46].

Concerns over net energy balance, potable water use, environmental hazards, and processing technologies call into question the potential for terrestrial biomass such as cereals crops and lignocellulose biomass to provide efficient sustainable answers to future food and energy challenges [19].

At the same time, an expanding body of evidence has demonstrated that marine macroalgae can provide a sustainable alternative source of biomass for food, feeds, fuel and chemicals generation [3, 33, 57, 58]. Lehahn et al. [34] identified the "potential reserves" analogue of near offshore macroalgae for biofuels. Their calculation suggests that near-future technologically and economically deployable areas, associated with up to 100 m water depth, and 400 Km distance from the shore, can provide 10^9 dry weight (DW) ton per year, which is equivalent to 18 EJ[1] of energy.

There are several properties of macroalgae, which make them attractive feedstock for biofuels and high value chemicals (Table 1). First, macroalgae grow faster than terrestrial plants [8, 18, 27]. Second, macroalgae do not occupy arable land and do not consume fresh water [16], if cultivated offshore; thus they do not compete with traditional food agriculture [10, 21]. Third, macroalgae normally contain no or less lignin, eliminating the energy intensive lignin removal step in pre-treatment processes

[1] Exajoule.

Table 1 Advantages and limitations of macroalgae feedstock for biofuels (comparing to 1st generation biofuels)

Advantages	Limitations
Off-shore cultivation – no competition for land and potable water	Higher quantities of ash/salt, sulfur, and nitrogen
Rapid growth	Higher bromine and iodine content
Low content of lignin	Higher water content
Uptake of inorganic nutrients	Higher, metal and halogens content
Higher photosynthetic efficiency	Lower heating values (same as for wood chips)
Potential for long lines of co-production	

[4, 9]. The high carbohydrate content of macroalgae also makes them suitable for bioconversion into platform chemicals and biofuels molecules such as methane [36], hydrogen [47], syngas [52], ethanol [27], n-butanol [44], 2,3-butanadiol [37], etc.

Yet, the economic analyses of macroalgae as energy feedstock are scarce [20, 30]. Roesijadi et al. [46] highlighted the need for advances throughout the supply chain, and called for a detailed assessment of environmental resources, cultivation and harvesting technology, conversion to fuels and high value chemicals, connectivity with existing energy supply chains, and the associated economic and life cycle analyses in order to facilitate evaluation of this potentially important biomass resource. Moreover, no decision making frameworks addressing the economic challenges of introducing and commercializing these technologies are identified in the literature. The literature suggests that more quantitative understanding of the economics is essential for the development of the industry [28, 49]. The aim of this study is to investigate macroalgae utilization practices in present, and to characterize the key challenges for profitable macroalgae-based industry in the future. The study investigates the opportunities, advantages, limitations and other issues encountered to this emerging industry. The paper is structured as follows. In the next section macroalgal biomass as potential natural resource for a variety of outputs is presented. Section 3 compares five representative research studies to illustrate current state of the art of the macroalgae technological feasibility. Stylized profitability analysis is presented in Sect. 4. To conclude the paper, the key challenges of macroalgae utilization are discussed in Sect. 5.

2 Macroalgae Industry

Macroalgae have been harvested throughout the world as a food source and as a commodity for the production of hydrocolloids for centuries. Increasingly, seaweed is cultivated rather than collected from the wild. According to FAO statistics, the share of wild seaweed in global seaweed production fell from 28% in 1980 to 4.5% in 2010. This declining share reflects both the increased volume of cultivated seaweed and an absolute decrease in wild seaweed tonnage [56]. Currently the industry of macroalgae cultivation is mainly concentrated in Asia [24].

2.1 Seaweed Chemical Composition

Chemical composition of macroalgae species is significantly different from terrestrial plants (Table 2). They include lower contents of carbon, hydrogen, and oxygen and higher contents of nitrogen and sulfur than that of land-based, lignocellulosic biomass. Macroalgae nearly absent lignin [30], as opposed to ∼ 16% lignin in the case of corn stover [20]. And, the seaweed grows more rapidly than terrestrial crops due to higher efficiency of photosynthesis.

These characteristics of macroalgae have economic implications on both private and external costs and benefits of macroalgae-based feedstock utilization.

2.2 Energy

Over the years, many researchers have examined biofuel production from various types of macroalgae. Conversion factors of green seaweed to energy products, as reported in the literature, are summarized in Table 3.

Importantly, macro alga is a promising source for renewable energy production since it can fix the greenhouse gas (CO2) by photosynthesis [8]. The average photosynthetic efficiency is about 5% [2] - much higher than that of terrestrial biomass (1.8-2.2). Dissolved inorganic nutrients like nitrogen, phosphorous and carbon are taken up by macroalgae, helping to alleviate eutrophication in seas and oceans [17].

Table 2 Potential of green seaweed biorefineries (gr per Kg of DW)

Protein fraction	262
Fatty Acids	21.1
Glucose	113
Rhamnose	90
Xylose	29
Galactose	70
Ash	173

Table 3 Green macroalgae-based energy potential

Biomass DW derived product	Conversion factor	Reference
Ethanol [gm^{-2}]	0.14 (0.03–0.23)	Nikolaisen, et al. 2008
Buthanol	0.03–0.06	[57]
Ethanol	0.03–0.23	[57]
Acetone	0.01–0.02	[44, 57]
Methane[m3/tonDW]	10–96	[3]
Protein[gm^{-2}]	0.18	[1, 57]
Energy [KJm^{-2}]	19	[59]

(*Source Lehahn et al.*[34])

Table 4 Macroalgae-based protein potential

Market segments	Products segments	Total available market ($'B)	Serviceable available market ($'B)	Serviceable obtainable market ($'M)
Pharma	Gene expression systems	$5B	$2.5B	$50M
Cosmetics	Skin anti-aging and anticCellulite	$100B	$1B	$10M
Food	Macro algae food	$6B	$1B	$6M
	Protein ingredients	$18B	$7.2B	$30M
	Meat substitute	$3.6B	$1B	$20M
	Carbohydrates ingredients	$500B	$1B	$20M
Feed	Aquaculture feed	$5.1B	$700M	$6M
	Protein animal feed	$70B	$30B	$10M
Agriculture	Fertilizers	$175B	$240M	$3M

2.3 *Food and Proteins*

Apart from biofuel, macroalgae have a potential for additional end uses [20]. Macroalgae can be used to co-produce food and high value chemicals[2] (Table 4). Hochman and Zilberman [23] report that the food industry of macroalgae is estimated to generate $5 billion a year, a further $600 million is estimated to have been generated from hydrocolloids extracted from the cell wall of the macroalgae at an average value of about $10,900 a ton.

Protein market is assessed as $100B with a growing rate of 4.5%. From it protein feed market is about $70B and growing 5% per year (Table 3). The protein food ingredients market is $18B, while in the USA alone it is $4.5B with a growing rate of 8–9%. Plant protein ingredient marker is assessed globally at $5.4B, with $2B in the USA and a growing rate of 8%. The main factors for plant protein market growth are industrial farming (20% growth in 5 years in USA), population growth 1.3%, increasing nutritional and food safety requirements, and consumers' health-consciousness. An increase of 29% in high protein products was reported (DuPont, USDA, Martec, Euromonitor). Moreover, protein from macroalgae can supplement the soybean. Today, owing to its high protein content, the soybean is probably the single most important protein crop in the world. From 2005 to 2010 soy USA protein market doubled. The demand for plant proteins is expected to continue to grow and so the environmental pressure due to the industrial agriculture and growing vegetarian

[2] Carrageenan, mannitol, agar, laminarin, mannan, ulvan, fucoidin, and alginate, carbohydrates, (mannitol has a lower calorific value, and has been found to be effective as a sweetener in various food product and pharmaceutics). Extracted algin quickly absorbs water (200–300times its own weight) and thus is effective as an additive in dehydrated products, as well as the paper and textile industries. It has also found use as a food thickener and stabilizer.

population. In 2005 the vegetarian food market reached $1.2B sales in the US alone. Meat substitutes sales reached $326M in the US and $2B in Europe at the end of 2009.

The use of macroalgae as a potential source of high value chemicals and in the rapeutic purpose has engrossed its commercial interest on macroalgae. For example, the most diversely used macroalgae derivative with substantial worldwide sales is Agar. The highest-value derivative of agar is called agarose and is used in a microbiological genetic-engineering application. Furthermore, macroalgae have shown to provide a rich source of natural bioactive compounds with antiviral, antifungal, antibacterial, antioxidant, anti-inflammatory, hypercholesterolemia and hypolipidemic and anti-neoplasteic properties [51].

Another example is Carrageenan - a gelling agent extracted from red seaweeds. It can be used as an emulsifier, a binder, or for suspension and stabilization in a remarkably wide range of products in the food processing, pharmaceutical and cosmetic industries. As an approved food additive, carrageenan is used worldwide to enhance dairy and meat products; it also has a variety of applications ranging from toothpaste to pet food. According to FAO statistics, world carrageenan seaweed farming production increased from less than 1 million wet tonnes in 2000 to 5.6 million wet tonnes in 2010, with the corresponding farmgate value increasing from USD72 million to USD1.4 billion [56].[3]

In the next section we present the schematic structure of off-shore macroalgae cultivation and biorefinery, in order to define key drivers for the profitable production process.

2.4 Production Structure of Macroalgae

Macroalgae based production consists of several processes, each may be affected by a specific technology or input. Figure 1 schematically outlines the key milestones of macroalgae to biochemicals' production. Cultivation, harvest and transportation are plant based decisions, whereas pretreatment and conversion are subject to refining technology.

Anaerobic digestion, fermentation, transesterification, liquefaction and pyrolysis can convert algal biomass into proteins and sugars that can result into food, chemicals and biofuels. At each stage of the production process the investor should decide between various options that ultimately affect the irreversible (sunk) and variable costs of the production, the productivity, and the output, therefore affecting the total profitability. The configuration of baseline production characteristics should be defined according to available resources and best technologies (e.g. [25, 30, 56]. In addition to the production cost, the value of seaweed products when reaching

[3] Valderrama et al. [56] present an overview of trade trends of carrageenan in the 2000s.

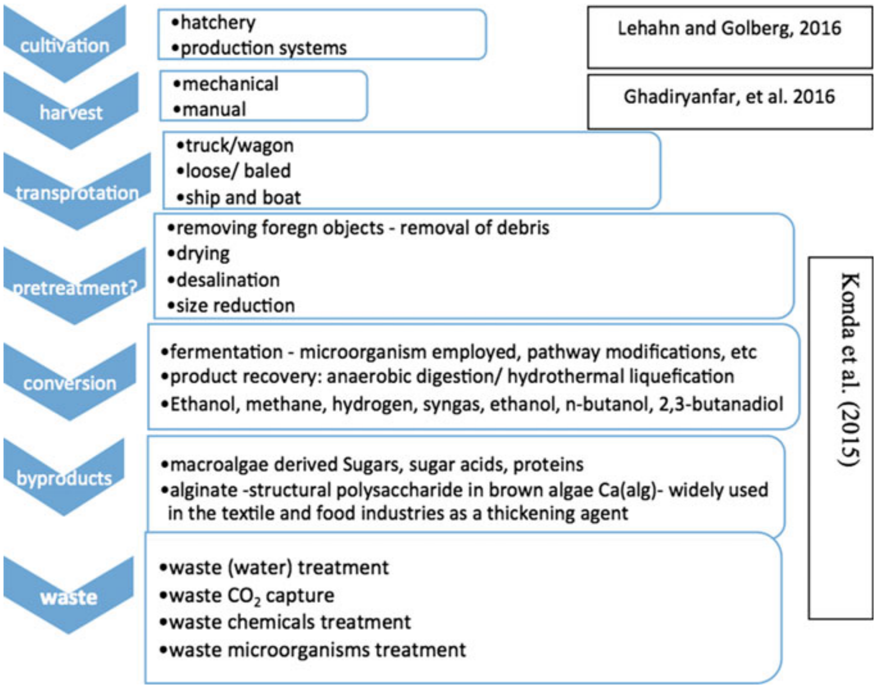

cultivation
•hatchery
•production systems
Lehahn and Golberg, 2016
harvest
•mechanical
•manual
Ghadiryanfar, et al. 2016
transprotation
•truck/wagon
•loose/ baled
•ship and boat
pretreatment?
•removing foregn objects - removal of debris
•drying
•desalination
•size reduction
Konda et al. (2015)
conversion
•fermentation - microorganism employed, pathway modifications, etc
•product recovery: anaerobic digestion/ hydrothermal liquefication
•Ethanol, methane, hydrogen, syngas, ethanol, n-butanol, 2,3-butanadiol
byproducts
•macroalgae derived Sugars, sugar acids, proteins
•alginate -structural polysaccharide in brown algae Ca(alg)- widely used in the textile and food industries as a thickening agent
waste
•waste (water) treatment
•waste CO_2 capture
•waste chemicals treatment
•waste microorganisms treatment

Fig. 1 Overview of macroalgae cultivation and processing

end users may also reflect the expenses on research and development (R&D), formulation, marketing, etc. [56]. Specific information on these aspects is generally lacking.[4]

3 Case-Studies of Macroalgae Feasibility Analysis

In order to illustrate current state of the art of the macroalgae technological feasibility we compare five recent representative research studies summarized in Table 5. Cases I and II focus on techno-economic (TEA) analysis of cultivation of macroalgae, Case III provides a TEA of the biorefinery for macroalagae-based biofuels, Case IV focuses on the cost-benefit analysis (CBA) of an experiment-based cultivation of macroalgae as well as the biorefinery for macroalagae-based biofuels, while Case V presents the life cycle analysis (LCA) of the system that includes cultivation and processing.

[4] Composed based on Konda et al. [30] and Ghadiryanfar et al. [20].

Case I – TEA Cultivation [11]

This report, prepared in the framework of the European project EnAlgae, attempts to supply a transparent cost - revenue estimation for Laminaria digitate (brown seaweed) cultivation in North Western Europe (UK, Ireland, France and the Netherlands). The cultivation process is split up in the cultivation of plant material (culture strings with juvenile sporophytes) in a hatchery for 3–5 months and the on-sea production of seaweed biomass, where after growing out for 5–6 months the biomass is harvested in the spring. The detailed cost break includes among others sea aquaculture license. The study reaches a rather high price for macroalgae at the aquafarm gate required to cover production costs.

Case II – Cultivation [56]

This report, prepared for FAO, performs a cost-benefit analysis of carrageenan seaweed farming in four developing countries, accounting for about 90% of world cultivation of carrageenan seaweed in 2010: Indonesia, the Philippines, Solomon Islands, and the United Republic of Tanzania. It presents data on costs and revenues from actual cultivation of Kappaphycus and Eucheuma (red seaweed species) and supply chains in place. The cultivation is observed most of the year with 4 to 8 cycles of 45 days. Even though most of the 23 case studies revealed profitability, defining conditions for profitable production proved difficult. No distinct patterns in the productivity of different farming systems are detected, neither in terms of production per unit of cultivation line, nor in terms of production per unit of farming area. The authors claim that the direct comparison of the productivity of two farming systems may reflect mostly the differences in their farm locations (e.g. temperature, weather condition, and water quality) that affect the growth rate of seaweed and the number of growing cycles (as two primary factors determining the productivity).

Case III – TEA of Biorefinery [30]

The research presents a TEA of a simulated macroalgae biorefinery for fermentation-derived sugars, and specifically ethanol with co-production of alginate. This study does not assess the cultivation process. Instead, it assumes (rather low) feedstock price of $50–100/MT of DW for brown seaweed Saccharina latissima. The laboratory-based conversion technology is scaled-up to simulate the fermentation-derived products from macroalgae at an industrial-scale facility with 2000 MT/day dry biomass processing capacity. The cost structure is largely based on the study by the National Renewable Energy Laboratory (NREL) on the production of ethanol from corn stover [25], modified for the processing of macroalgae and its products. Results suggest the minimum ethanol selling price (MESP) in the range of $3.6–8.5/gal. For production of chemicals, sugar prices were in the range of 21–47/lb or 16–40/lb with macroalgae priced at $100/MT and $50/MT, respectively.

Case IV – CBA of Cultivation and Biorefinery [31, 32]

This study performs a CBA for bioethanol production using biomass of Ulva rigida (green seaweed species), co-cultured with fish in an intensive offshore aquaculture unit. This report takes into consideration offshore seaweed cultivation during summer and uses an ethanol production technology that is devoid of pre-treatments. Co-production of ethanol and Dried Distillers Grains with Solubles (DDGS) is considered. Growth yields in the off shore experiments in Israel are extrapolated to project large-scale production volumes. The economic analysis is performed using costs from studies by NREL [25]. The costs associated with the by-product subprocess are based on figures of dry-grind corn processing. For profit calculation, the study assumes prices for ethanol and DDGS according to the U.S. Department of Agriculture (USDA), (79/dry ton of U. rigida. and -630/dry ton of U. rigida respectively) and asks what production volume reaches profitability. The authors claim that only large scale production shows economic viability.

Case V - LCA of Cultivation and Biorefinery [48]

The study performs a comparative life cycle assessment of the offshore seaweed cultivation for the production of biorefinery feedstock. The biomass is converted into three products: bioethanol, liquid fertilizer and protein-rich ingredient for fish feed. The system represents Danish conditions with Laminaria digitata (brown seaweed species) average productivity of 10 Mg WW/ha and harvested in summer. There are no costs reported, but the authors adapt the design of cornstover bioethanol production by NREL [25] to model energy consumption in the industrial scale seaweed based biorefinery with bioethanol production using separate hydrolysis and fermentation. The results of this study show that the base case provides a net reduction in climate change factors. However, for the base case the research reports an increase in human toxicity that is seven times greater than the system can deal. The study indicates that the hotspot in the value chain is the biomass productivity.

Table 5 synthesizes the representative studies from developed and developing countries that employ different cultivation and conversion technologies, and allows to draw several general conclusions:

1. Not all macroalgae are the same: Various macroalgae species (different colors) allow for different outputs. A critical decision in the offshore biomass production for biofuels is the species choice [17]. For example, different macroalgal species could be chosen for their production of low-cost fuel in combination with high value compounds and/or bioremediation applications, where an excess of nutrients can be converted in biomass for harvest and economic goods. Thus the entrepreneur needs first to decide what outputs to produce and then what seaweed species to use as feedstock.
2. Feedstock production uncertainties: Even though production uncertainties are inherent in agriculture, farmers growing cellulosic energy crops face significant revenue uncertainty. Important constrains include light, temperature, nutrients, current velocity, and also the capability to resist the harsh conditions such as high waves and extreme currents in offshore waters. Nitrogen has often been indicated as the primary limiting factor for seaweed growth; however, phosphorus may also limit production in some systems [17]. Other environmental factors negatively affecting the performance of seaweed farming include grazing by fish or other organisms and rising sea temperatures, which could slow seaweed growth [26]. Literature reports a wide range of growth rates for macroalgae. In the FAO study (Case II) the growth rate varies between 0.2 to 10.86% per day for red seaweed, while Korzen et al. [31, 32] (Case IV) reach 15% of average daily growth for green seaweed.

Table 5 Representative recent studies on macroalgae utilization[†]

Reference Parameter	**Case I** (Dijk and Schoot 2015) **ENAlgae Cultivation**	**Case II** (Valderrama eds. 2013) **FAO Cultivation**	**Case III** (Konda, et al. 2015) **Bio-Refinery**	**Case IV** (Korzen, Peled, et al. 2015) **Cultivation and Bio refinery**	**Case V** (Seghetta, et al. 2016) **LCA**	
Macroalgae (Seaweed) type	Brown: Laminaria Digitata	Red: Kappaphycus And Eucheuma	Brown: Saccharina Latissima	Green Ulva Rigida	Brown: Laminaria Digitata	Brown: Saccharina Latissima
Country of experiment	Ireland, UK, France, Holland	Philippines, Indonesia, Tanzania, India, Mexico, Solomon Islands	Simulation	Israel	Denmark	
Output under study	Seaweed	Carrageenan, Gracilaria (primary raw materials for agar) and nori	co-production ethanol and alginate	DDGs fish feed and ethanol	co-production: ethanol, liquid fertilizer, fish feed	
Growth Rate (Average)	10 kg WW/m long-line	0.2%-10.86% daily average growth rate WW		15% daily average growth rate WW	average productivity of 10 Mg WW/ha	average productivity of 1.5 Mg WW/ha
Season for harvesting	Spring	4 to 8 cycles a year	NA	April-October	Summer	
DW/WW Ratio				1/9	1/3	1/6
Yield in biorefinery			0.15 ethanol	0.12 ethanol; 0.6 DDGs	0.005	0.13
Source of Costs	simulated	actual	NREL - simulated	NREL - simulated	NREL - simulated	
Conversion Technology	not assessed	not assessed	no pre-treatment, hedrolysis and fermentation	no pre-treatment, single step for the release of glucose, simultaneous fermentation to ethanol	bioethanol production using separate hydrolysis and fermentation (SHF);	
Price to Farmer ($/Ton DW)	6921[‡]P	500-1000	100 (21-120)	630		
R- interest rate	r - 5.5%, insurance 0.5%		irr 10%	5%		
Reported Profitability Indicator	Cost per kg WW to selling price	Observed productivity, efficiency and (mostly positive) profitability	MSP $6.510.5/gal ethanol; MSP $ 3.1 alginate (?)	NPV profitable at large scale	LCA ,externalities, no profitability reported; potential for carbon sink, but also for increase in human toxicity (cancer)	
Price of Output	~ $10,000/ton DW	$2500/ton average	MESP $8.5/gal ethanol, $3.1/kg alginate	€79/DW ton of U. rigida for ethanol and 630/DW ton of U. rigida for DDGS		

† The color of each column corresponds the color of the seaweed species under study.

‡Personal calculation based on the reported price in €/ton WW.

3. Processing technology uncertainties: The biorefinery yields present a wide range as well. The upper value can be ten times larger than the lower one (Tables 2 and 5), significantly affecting the potential profitability of the process.
4. Variability of DW/WW ratio: Dry weight to wet (fresh) weight ratio of macroalgae is another parameter which values vary significantly (from 1/9 to 1/3, Table 5).
5. Price uncertainties: should be analyzed in several aspects – price uncertainties that faces the farmer, price uncertainties of feedstock for biorefinery, and the price uncertainty of competitive outputs (backstop technology). A seaweed industry that contains many small-scale pricetakers is especially prone to boom-bust cycles. For example, the strong demand from China drove the price of dry cottonii in the Philippines from USD900/tonne in 2007 to almost USD3 000/tonne in 2008 causing the Philippines production to double from 1.5 million tonnes (wet weight) in 2007 to 3.3 million tonnes in 2008. The "seaweed rush" lasted only one year – the price dropped to USD1 300/tonne in 2009 [26]. Generally, when strong demand for dry seaweeds drives up the price, seaweed farmers tend to increase their planting efforts and/or harvest immature crops. However, if the price is low, seaweed farmers tend to reduce production, which creates sourcing difficulties for the local processors. On the other hand, processors would tend to reduce demand as prices rise by substituting cheaper alternatives [38]. A likely result would then be supply exceeding demand and consequently a collapse in price.
6. The price and cost assumptions in the academic literature should be treated cautiously and verified against actual data, as prices vary over time and may experience sudden picks or drops. For instance, actual or assumed price to farmer in Table 5 varies from $50/MT of DW to $10,000/MT.
7. Production functions: The studies' effort to evaluate future costs of the process that is currently available mostly in small (lab) scale is remarkable and should not be underestimated. However, the studies mostly lack (or do not report) a structured production function that leads to a cost function. The common assumption is a linear approximation.
8. Single cost sources: All the studies that assessed biorefinery used the conveniently available calculation module on large-scale ethanol production from cornstove by NREL [25]. Indeed, the popularity of this research signals that more up-to-date studies with a transparent open-source tool would increase our understanding of the economic viability of the novel biorefinery technologies.
9. Developed versus developing: In contrast to aquaculture in the developed countries, carrageenan seaweed farming in Asia has minimum capital and technological requirements and, as such, produces feedstock at competitive prices.
10. Supply chain: there are established supply chains for seaweed used for food production [56]. Supply chains for biorefinery processing are still to be developed. Contract uncertainties may occur due to asymmetric information [12]. That is, the innovator may not observe the ability of and effort being devoted by the contracted supplier.

11. Value of environmental amenities: Even though intuitively macroalgae based biofuel is cleaner than fossil fuels, the environmental advantages still require more investigation. Seaweeds could improve the benthic ecosystem, and sequester carbon, thereby offering the potential for carbon credits. Seaweed grown on rafts can also become an attractive haven for fish. Other positive environmental externalities of seaweed farming include an alleged positive attitude towards conservation of local marine habitats, and some evidence that overexploitation of the fisheries has been reduced in some countries, because farmers have less time or inclination to fish. However, negative externalities should not be overlooked. For instance, disease is a major problem in the cultivation process, which not only discourages farmers but also contributes to supply uncertainty for processors. Ice-ice disease is a common disease that affects carrageenan seaweed farming worldwide. Primarily because of perennial ice-ice outbreaks, cottonii cultivation in Zanzibar (the United Republic of Tanzania) declined from over 1000 tonnes in 2001 to almost zero in 2008 [56]. In addition, introduced seaweed that do not become viable culture species could turn into an environmental nuisance [41].

4 Profitability of Macroalgae Cultivation

In order to demonstrate the general conclusions outlined in Sect. 3 above, we provide the profitability analysis for Case II – cultivation and food production (carrageenan) in major macroalgae farming countries and compare the results with the economic indicators of the seaweed cultivation in Europe – Case I.

4.1 The Analytical Framework

The economic performance of seaweed farming is determined by its economic costs and benefits. The main economic costs include capital, material inputs and labour. The economic benefits can be measured by the revenue and cash flow generated by seaweed production. Profit is an indicator of the net benefit, which measures tradeoffs between benefits and costs. Various performance indicators (e.g. productivity, efficiency and profitability) are used to compare the economic costs and benefits.

In order to determine the economic feasibility of macroalgae -based production process, an assessment of Net Present Value (NPV) using operational revenues and costs, as well as capital costs, which are all linked to varying production volumes, is performed.

Let all expenses at point t be denoted with $C(t)$ and all returns with $B(t)$. An investment into macroalgae utilization process is profitable if it provides a positive NPV, meaning the discounted sum of expenses and returns is positive:

$$NPV = \sum_{t=0}^{T} (B(t) - C(t))\, q^{-t}, \tag{1}$$

where $q^{-t} = (1+i)^{-t}$ is the discount factor, i is the annual rate of return. At the breakeven point (zero profit condition), total costs should be equal to total revenues. In terms of Eq. (1) it means:

$$\sum_{t=0}^{T} B(t)q^{-t} = \sum_{t=0}^{T} C(t)q^{-t}. \tag{2}$$

Minimal selling price (MSP) or breakeven sale price is an additional economic indicator that represents the zero-profit threshold i.e. the price that covers the costs per unit of production. It is estimated as the annual cost of capital per unit of production plus the variable cost per unit. In Sect. 4.2 we use the economic tools to characterize the profitability conditions of different case studies.

4.2 Profitability of Carrageenan Seaweed

What can we learn from existing seaweed cultivation practices? In order to answer this question, we analyze the data from 13 case studies from largest carrageenan seaweed farming countries: The Philippines, Indonesia, Tanzania and Solomon Islands. The four case-study countries accounted for about 90% of world cultivation of carrageenan seaweed in 2010 [56]. Common characteristics of red algae cultivation in these developing countries are summarized in Table 6. The cultivation of fresh seaweed is usually conducted by a number of small-scale, independent seaweed growers. Various cultivation practices are in place and can be typed into: off-bottom, floating ramp and floating line. The lifespan of cultivation farm commonly lasts from 2 to 5 years. Fresh seaweeds decompose quickly after harvest. Sun-drying remains the main (if not the only) option in practice. The industry standard for the maximum moisture content of dry cottonii is 38–40%. Post-harvest treatment is usually done by seaweed growers.

The physical capital needed for carrageenan seaweed farming usually includes farming systems, vessels, shelters, drying facilities, and miscellaneous equipment or tools. Figure 2 presents capital efficiency measured in terms of initial investment in US$ per km of cultivation line and yield reported in each of the case studies as annual productivity of cultivation line in tonne of DW per km.

Evidently, more capital does not insure a higher yield. Based on the case studies from developing countries, Fig. 2 provides a rough approximation for marginal productivity of capital reflecting its diminishing nature. The capital investment per km of cultivation line is lower in case studies with higher production. Similarly, no cultivation technique (off-bottom, floating raft or line, etc.) is observed to be the most

Table 6 Kappaphycus farming in the case-study countries

Country	Notes	Farming system	Cultivation line (km)	Farm area (ha)	Annual production tone DW/year	Productivity tone/year/km
Indonesia	representative small-scale nuclear family farm	Floating raft	6	–	6.6	1.1
	representative large-scale leader farm	Floating raft	30	–	33	1.1
Philippines	Six representative farms using different farming systems and/ or in different locations	Off-bottom	1.8	–	2.143	1.8
		Off-botton	1.62	–	0.9	1.62
		Floating line	2.7	–	8.57	2.7
		Floating line	1.8	–	2.75	1.8
		Floating raft	–	0.05	2.85	–
		Floating line	–	0.27	8.5	–
Tanzania	Two representative farms using different farming	Off-bottom	0.3	–	0.662	0.3
		Floating line	0.324	-	0.806	0.32
Solomon Islands	representative farms based on field survey	Off-bottom	4	–	17.4	4
			4	21.7	4	
			2.4	9.2	2.4	

(Based on Valderrama Eds. [56])

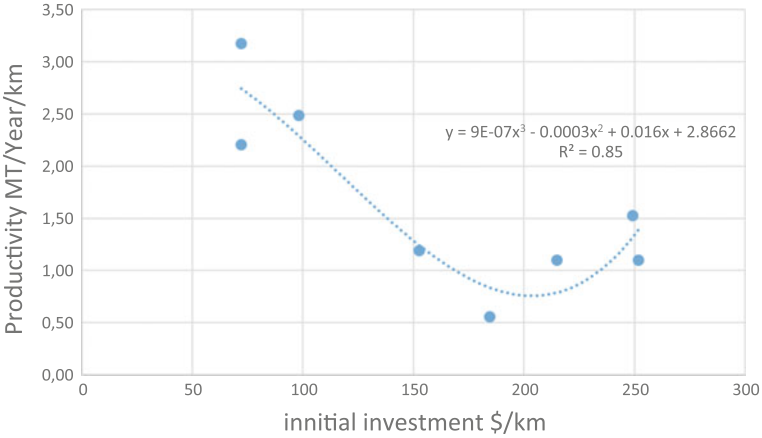


Fig. 2 Capital efficiency: initial investment and yield per km of cultivation line

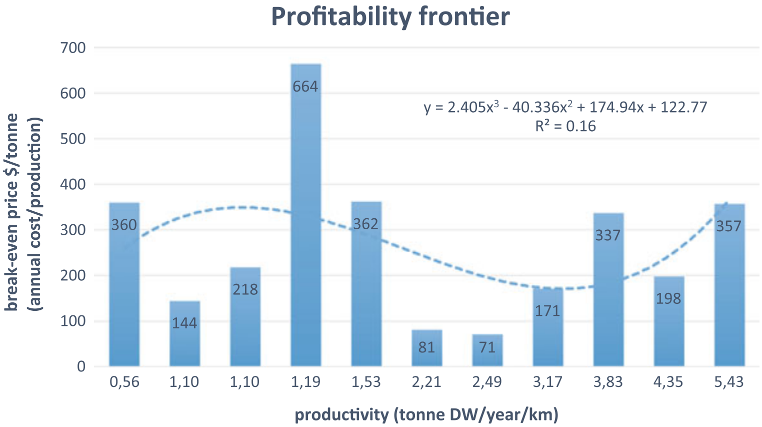


Fig. 3 Profitability frontier

efficient. Possible explanation is that in addition to capital and cultivation technique there are other major growth affecting factors, such as seasonality and inclement weather. Nevertheless, we are in early stages of investigation of the technology and much more information is needed to reach solid conclusions.

Figure 3 presents the breakeven price equivalent to MSP and defined in Eq. (2), and the annual productivity per km of cultivation line. Higher break-even price is addressed to more sophisticated and/or commercialized farms. In this case the breakeven price is the minimal farm gate price for profitable cultivation. Accordingly, the trend line is a stylized profitability frontier indicating of the lower limit for market prices for unprocessed seaweed. All the case studies in the developing countries had positive profits (Fig. 4), ranging from USD89 per tonne of dried seaweed to

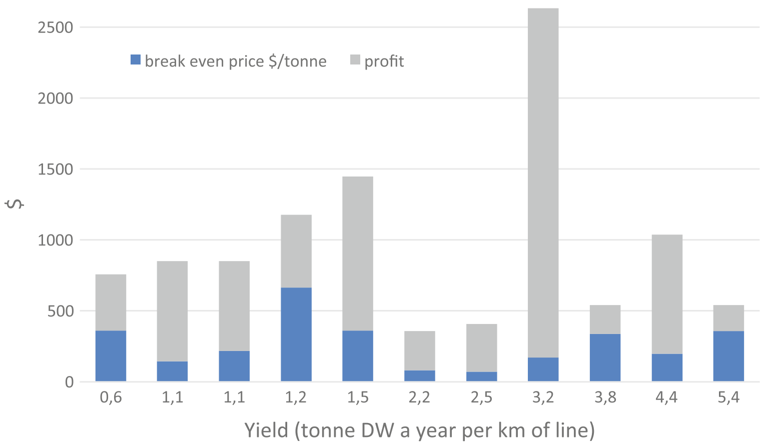


Fig. 4 Yield, break-even price and the profit

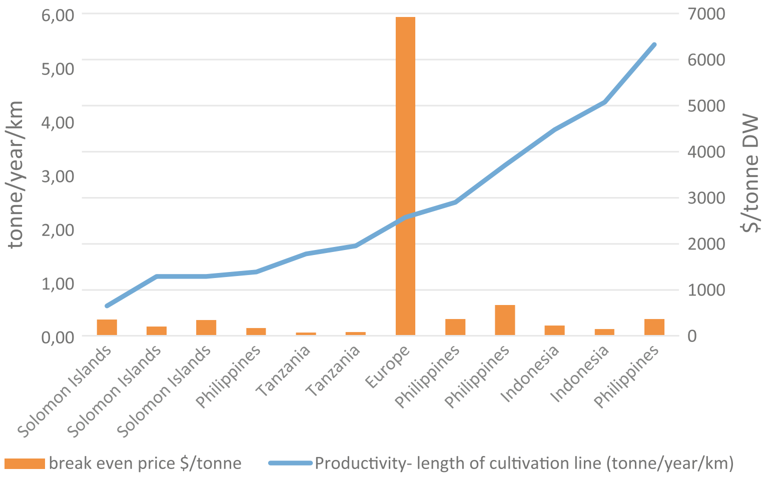


Fig. 5 Yield and break even price in the case studies in developing countries and Northern Europe

USD842/tonne. Macroalgae price is the key factor affecting profit. The profit margin (i.e. the ratio of profit to farm revenue) of most of the cases exceeded 50%.

To compare, the yield reported in the Case I – EnAlgae project for the North Western Europe [11], is within the middle of the range of the Case II studies (Fig. 5). However, the break-even price, as we calculated based on the EnAlgae project, is on average 10 times higher than the annual costs in the case studies collected in the developing countries. Given cost differences, developing countries may be able to sustain algae production systems with lower yields.

The difference in costs is explained by a short production cycle, low capital requirement, and relatively simple farming technology in developing countries (Case

II) versus the North-Western Europe (Case I). Macroalgae cultivation and post-harvest treatment as reported for North-Western Europe are labour – intensive activities entailing relatively high amounts of initial capital and laboratory costs.

5 Summary and Discussion

In this work we discuss the economic opportunities and challenges of macroalgae utilization. The study outlines the state of the art of technological and economic abilities of macroalgae cultivation and conversion. The focus on macroalgae is driven by the fact that being cultivated off-shore, they do not compete for scarce land and potable water. In addition, recent developments in bio-refinery show the potential to produce not only food and coloring, but also sugars for biofuels, proteins, and high value chemicals. Evidently, carrageenan seaweed farming, has evolved into a successful commercial endeavor in a number of tropical countries endowed with clear, unpolluted intertidal environments and protected beach locations.

Nevertheless, several major challenges should be taken into consideration for successful macroalgae economy. First, the rate of macroalgae growth and the conversion factors – two key parameters in productivity- show a wide range of values and therefore have a major effect on cost effectiveness of the technology. Macroalgae growth depends on saturation kinetics by light intensity, ambient dissolved inorganic nutrient concentrations and temperature [7]. Cultivation uncertainty is exacerbated by stochastic weather, seasonal variability between regions, within years and between years. The biomass productivity is the main constraint against being competitive with other energy and protein producing technologies [48].

Previous studies suggested to combine macroalgae cultivation with other sea related economic activities [5, 6, 29, 31, 32, 39]. Co-management with other off-shore systems like wind farms and fisheries to increase economic and environmental benefits, and to diversify the revenue sources should be considered.

Another way to diversify revenue is the co-production in the stage of feedstock conversion (to e.g. biofuels and food). The variation in shares of co-products between, for example, butanol-acetone, ethanol and methane as well as protein, may affect substantially the net benefits of the production process. More research on the key aspects of co-production that leads to increased profitability of biorefinery is crucial.

Next, investments in macroalgae utilization are risky not only due to the uncertainty in feedstock cultivation, but also in processing technology, contracting, and demand. Considering uncertainty is most pertinent when a new processing technology, such as new biofuel refining technology, is invented. Design of a sustainable biorefinery, which will generate sustainable food, fuels and chemicals is a complex task and is largely influenced by local raw material supplies, advances in multiple technologies and socio-economic conditions [15]. In addition, comprehensive scientific studies on the question whether the novel bio-refinery can increase the yield by the order of 10 in a rigid manner to assure profitable process should be undertaken.

Decisions about the scale of operation and the division of supply of inputs between in-house and external operations are key in the design of a basic supply chain [12]. These decisions are affected by the investors financial situation, the political and social system, the technology available, etc. The strategy about the capacity of feedstock plant as well as the refinery may change over time; the innovator may experiment by starting at small scale. Once the production system is established, the innovator may either expand operations or reach out to cooperatives to provide it with inputs. Therefore, more research on economies of scale in macroalgae cultivation and refinery is crucial for industry establishing.

Investments in production capacity or consumption infrastructure are also susceptible to market uncertainties from, for example, fluctuations in energy prices [45] and demand uncertainty that is often associated with new product introduction. Similarly to traditional crops, the price paid to seaweed farmers is determined in part by the complexity of the supply chain and partly by the quality of the macroalgae. But, crops destined for conversion into bio-fuels have prices determined in large part by the ethanol market, which is linked to the volatile gasoline market [54].

Energy crop price volatility is likely to be aggravated as ethanol shifts in and out of status as a cost-effective fuel substitute for gasoline, based on the relative prices of petroleum and corn grain, the leading current ethanol feedstock in the United States. More investigation on the impact of output price variability on technology adoption decisions is essential.

Price volatility is also compounded by the absence of relevant, reliable and timely production statistics and market intelligence. Unlike for some agricultural commodities such as coffee or tea, there are no organized markets to provide benchmarking international prices for seaweed [53]. Unavailability of reliable information is especially detrimental to uninformed seaweed farmers who are at the lowest end of the seaweed value chain and often forced to accept whatever price is offered.

Moreover, it is essential to identify the fuel that may provide higher value than the ethanol, as of now, maclaogae-based bioethanol cannot compete corn-ethanol or sugar cane based ethanol. To generalize, rather than competing with existing goods, the scientific challenge can be the investigation of the potential to utilize macroalgae for unique foods, high value chemicals and fuels.

Besides, the transparency and multidisciplinary interaction will increase the learning curve and will make macroalgae production more structured and efficient. Alternative specifications for biorefinery and cultivation processes in a transparent way would allow replication and induce the improvement of methodologies. The multidisciplinary effort is required for improving the knowledge of production and cost functions to lead to establishing of economic models.

Not less important is the uninvestigated effect the mass cultivation of offshore macroalgae might have on the environment. On the one hand, the transition from small to big scale macroalgae cultivation involves direct and external effects that may completely reshape the process. On the other hand, if macroalgae-based biofuel crowds-out the use of fossil fuels and crop-based bioethanol, it mediates the environmental externalities, as well as negative effects on agricultural supply and land use [60]. Further analysis on macroalgae external costs and benefits is required for an accurate policy intervention. The analysis on the technological prospects of macroalgae biorefinery should evaluate the social net benefit too. Consequently, the recommendation upon optimal fuel mix is to be based on social (vs. private) costs.

References

1. Abudabos, A.M., et al.: Nutritional value of green seaweed (Ulva lactuca) for broiler chickens. Italian J. Animal Sci. **12**(28) (2013)
2. Aresta, M., Dibenedetto, A., Carone, M., Colonna, T., Fragale, C.: Production of biodiesel from macroalgae by supercritical CO2 extraction and thermochemical liquefaction. Environ. Chem. Lett. **3**(3), 136–139 (2005)
3. Bruhn, A., Dahl, J., Nielsen, H.B., Nikolaisen, L., Rasmussen, M.B., Markager, S., Olesen, B., Arias, C., Jensen, P.D.: Bioenergy potential of Ulva lactuca: biomass yield, methane production and combustion. Bioresour. Technol. **102**(3), 2595–2604 (2011)
4. Bruton, T., Lyons, H., Lerat, Y., Stanley, M., BoRasmussen, M.: A review of the potential of marine algae as a source of biofuel in Ireland. (Sustainable Energy Ireland) (2009)
5. Buck, B., Krause, G., Michler-Cieluch, T., Brenner, M., Buchholz, C., Busch, J., Fisch, R., Geisen, M., Zielinski, O.: Meeting the quest for spatial efficiency: progress and prospects of extensive aquaculture within offshore wind farms. Helgoland Mar. Research **62**, 269–281 (2008)
6. Buck, B.H., Krause, G., Michler-Cieluch, T., Brenner, M., Buchholz, C.M., Busch, J., et al.: Meeting the quest for spatial efficiency: progress and prospects of extensive aquaculture within offshore wind farms. Helgoland Mar. Research **62**, 269–281 (2008)
7. Buschmann, A., Varela, D., Cifuentes, M., Hernndez-Gonzlez, M., Henrquez, L., Westermeier, R., Correa, J.: Experimental indoor cultivation of the carrageenophytic red alga Gigartina skottsbergii. Aquaculture **241**, 357–370 (2004)
8. Chen, H., Zhou, D., Luo, G., Zhang, S., Chen, J.: Macroalgae for biofuels production: progress and perspectives. Renew. Sustain. Energy Rev. **47**, 427–437 (2015)
9. Cho, Y., Kim, M., Kim, S.: Ethanol production from seaweed, enteromorpha intestinalis, by sepa-rate hydrolysis and fermentation (SHF) and simultaneous saccharifi- cation and fermentation (SSF) with saccharomyces cerevisiae. KSBB J. **28**, 366–371 (2013)
10. Daroch, M., Geng, S., Wang, G.: Recent advances in liquid biofuel production from algal feedstocks. Appl. Energy **102**, 1371–1381 (2013)
11. van Dijk, W., van der Schoot, J.R.: An Economic Model for Offshore Cultivation of Macroalgae. Public output report of the EnAlgae project, Swansea (2015)
12. Du, Xiaoxue, Liang Lu, Thomas Reardon, David Zilberman.: The Economics of Agricultural Supply Chain Design: A Portfolio Selection Approach. Am. J. Agr. Econ. **98** (5): 1377–1388. DuPont. n.d. Accessed Oct 2016. http://www.dupont.com/
13. Enriquez, J. 1998.: Genomics and the worlds economy. Science **281**: 925-926. EPA. n.d. Accessed Nov 2016. http://www.eia.gov/tools/faqs/faq.cfm?id=74&t=11
14. Euromonitor. n.d. *Time to Explore Algal Bioactives*. Accessed Oct 2016. http://blog.euromonitor.com/2013/10/time-to-explore-algal-bioactives.html
15. Fatih, D.M.: Biorefineries for biofuel upgrading: a critical review. Appl. Energy **86**, 151–161 (2009)
16. Fedoroff, N.V., Battisti, D.S., Beachy, R.N., Cooper, P.J.M., Fischhoff, D.A., Hodges, C.N., Knauf, V.C., et al.: Radically rethinking agriculture for the 21st century. Science **327**(5967), 833–834 (2010)
17. Fernand, F., Israel, A., Skjermo, J., Wichard, T., Timmermans, K.R., Golberg, A.: Offshore macroalgae biomass for bioenergy production: environmental aspects, technological achievements and challenges (in press)

18. Frost-Christensen, H., Sand-Jensen, K.: The quantum efficiency of photosynthesis in macroalgae and submerged angiosperms. Oecologia **91**(3), 377–384 (1992)
19. Gerbens-Leenes, W., Hoekstra, A.Y., van der Meer, T.H.: The water footprint of bioenergy. Proc. National Acad. Sci. **106**(25), 10219–10223 (2009)
20. Ghadiryanfar, M., Rosentrater, K.A., Keyhani, A., Omid, M.: A review of macroalgae production, with potential applications in biofuels and bioenergy. Renew. Sustain. Energy Rev. **54**, 473–481 (2016)
21. Goh, C.S., Lee, K.T.: A visionary and conceptual macroalgae-based third-generation bioethanol (TGB) biorefinery in Sabah, Malaysia as an underlay for renewable and sustainable development. Renew. Sustain. Energy Rev. **14**, 842–848 (2010)
22. Hannon, M., Gimpel, J., Tran, M., Rasala, B., Mayfield, S.: Biofuels from algae: challenges and potential. Biofuels **1**(5), 763–784 (2010)
23. Hochman, G., Zilberman, D.: Chapter 4 in Plants and Bio Energy. In: McCann, M., Buckeridge, M., Carpita, N. (eds.) Algae Farming and its Bio-products, pp. 49–64. Springer, New York (2014)
24. Hughes, A.D., Maeve, M.S., Black, K.D., Stanley, M.S.: Biogas from macroalgae: is it time to revisit the idea? Biotechnol. Biofuels **5**(1), (2012)
25. Humbird, D., Davis, R., Tao, L., Kinchin, C., Hsu, D., David, D., Aden, A.: Process design and economics for biochemical conversion of lignocellulosic biomass to ethanol. Natl. Renew. Energy Technol. 275-300 (2011)
26. Hurtado, A.Q.: Social and economic dimensions of carrageenan seaweed farming in the Philippines. In: Valderrama, D., Cai, J., Hishamunda, N., Ridler, N. (eds.) Social and Economic Dimensions of Carrageenan Seaweed Farming, pp. 91–113. FAO, Rome (2013)
27. John, R.P., Anisha, G.S., Nampoothiri, K.M., Pandey, A.: Micro and macroalgal biomass: a renewable source for bioethanol. Bioresour. Technol. **102**, 186–193 (2011)
28. Jones, C.S., Mayfield, S.P.: Algae biofuels: versatility for the future of bioenergy. Current Opinion Biotechnol. **23**(3), 346–351 (2012)
29. Jung, K., Lim, S.R., Kim, Y., Park, J.M.: Potentials of macroalgae as feedstocks for biorefinery. Bioresour. Technol. **135**, 182–190 (2013). (Pukyong National University)
30. Konda, N.V.S.N., Murthy, S.S., Simmons, B.A., Klein-Marcuschamer, D.: An investigation on the economic feasibility of macroalgae as a potential feedstock for biorefineries. Bioenergy Research **8**, 1046–1056 (2015)
31. Korzen, L., Yoav, P., Shiri Zemah, S., Mordechai, S., Aharon, G., Avigdor, A., Alvaro, I.: An economic analysis of bioethanol production from the marine macroalga Ulva (Chlorophyta). Technology, vol. 3(2), World Scientific Publishing Co (2015)
32. Korzen, L., Peled, Y., Shiri Zemah, S., Mordechai, S., Aharon, G.: An economic analysis of bioethanol production from the marine macroalga Ulva (Chlorophyta). Technology, vol.3 (2, 3), pp. 114–118. World Scientific Publishing Co (2015)
33. Kraan, S.: Mass-cultivation of carbohydrate rich macroalgae, a possible solution for sustainable biofuel production. Mitig. Adapt. Strateg. Global Chang. **18**, 27–46 (2013)
34. Lehahn, Y., Kapilkumar,.N.I., Alexander, G.: Global potential of offshore and shallow waters macroalgal biorefineries to provide for food, chemicals and energy: feasibility and sustainability. Algal Research **17**, 150–160 (2016)
35. Martec. n.d.: http://www.martecgroup.com/expertise/energy/. Accessed Oct 2016
36. Matsui, J.T., Amano, T., Koike, Y., Saiganji, A., Saito, H.: 2006. Methane Fermentation of Seaweed Biomass. American institute of chemical engineers (2006)
37. Mazumdar, S., Lee, J., Oh, M.-K.: Microbial production of 2,3 butanediol from seaweed hydrolysate using metabolically engineered Escherichia coli. Bioresour. Technol. **136**, 329–336 (2013)
38. McHugh, D.J.: The Seaweed Industry in the Pacific Islands. ACIAR Working Paper (2006)
39. Michler-Cieluch, T., Krause, G., Buck, B.H.: Reflections on integrating operation and maintenance activities of offshore wind farms and mariculture. Ocean and Coast. Manag. **52**, 57–68 (2009)

40. Nikolaisen, L., Jensen, P.D., Bech, K.S., Dahl, J., Busk, J., Brdsgaard, T., Rasmussen, M.B.: Energy Production from Marine Biomass (Ulva lactuca). Danish Technological Institute (2011)
41. Pickering, T., Skelton, P., Sulu, R.: Intentional introductions of commercially harvested alien seaweeds. Botanica Marina **50**, 338–350 (2007)
42. Pimentel, D. (ed.): Global Economic and Environmental Aspects of Biofuels. CRC Press, Boca Raton (2012)
43. Pimentel, D., Marcia H.P. (eds.): Food, Energy, and Society. CRC Press, Boca Raton (2007)
44. Potts, T., et al.: The production of butanol from Jamaica bay macro algae. Environ. Prog. Sustain. Energy **31**, 29–36 (2012)
45. Rajagopal, D., Sexton, S., Hochman, G., Zilberman, D.: Recent developments in renewable technologies: R&D investment in advanced biofuels. Annu. Review Resource Econ. **1**(1), 1–24 (2009)
46. Roesijadi, G., Jones, S., Snowden-Swan, L., Zhu, Y.: Macroalgae as a Biomass Feedstock: A Preliminary Analysis. PNNL 19944, Pacific Northwest National Laboratory, Richland (2010)
47. Sambusiti, C., Bellucci, M., Zabaniotou, A., Beneduce, L., Monlau, F.: Algae as promising feedstocks for fermentative biohydrogen production according to a biorefinery approach: A comprehensive review. Renew. Sustain. Energy Rev. **44**, 20–36 (2015)
48. Seghetta, M., Hou, X., Bastianoni, S., Bjerre, A.-B., Thomsen, M.: Life cycle assessment of macroalgal biorefinery for the production of ethanol, proteins and fertilizers - a step towards a regenerative bioeconomy. J. Clean. Prod. **137**, 1158–1169 (2016). doi:10.1016/j.jclepro.2016.07.195
49. Singh, A., Nigam, P.S., Murphy, J.D.: Mechanism and challenges in commercialisation of algal biofuels. Bioresour. Technol. **102**(1), 26–34 (2011)
50. Song, F., Jinhua, Z., Scott, M.S.: Switching to perennial energy crops under uncertainty and costly reversibility. Am. J. Agric. Econ. **93**(3), 768–783 (2011)
51. Suganya, T., Nagendra Gandhi, N., Renganathan, S.: Production of algal biodiesel from marine macroalgae Enteromorpha compressa by two step process: optimization and kinetic study. Bioresour. Technol. **128**, 392–400 (2013)
52. Suutari, M., Leskinen, E., Fagerstedt, K., Kuparinen, J., Kuuppo, P., Blomster, J.: Macroalgae in biofuel production. Phycol. Research **63**(1), 1–18 (2015)
53. Tinne, M., Preston, G., Tiroba, G.: Development of seaweed marketing and licensing arrangements. Technical report **1**, Project ST 98/009: Commercialisation of Seaweed Production in the Solomon Islands (2006)
54. Tyner, W.E.: The US ethanol and biofuels boom: its origins, current status, and future prospects. BioScience **58**(7), 646–653 (2008)
55. USDA. n.d.: https://naldc.nal.usda.gov/naldc/home.xhtml. Accessed Oct 2016
56. Valderrama, D., Cai, J., Hishamunda, N., Ridler, N. (eds.): Social and Economic Dimantions of Carrageenan Seaweed Farming. Fisheries and Agriculture Technical paper. FAO, Rome (2013)
57. van der Wal, H., Sperber, BLHM., Houweling-Tan, B., Bakker, RRC., Brandenburg, W., Lpez-Contreras, AM.: Production of acetone, butanol, and ethanol from biomass of the green seaweed Ulva lactuca.128 (2013)
58. Wargacki, A.J., Leonard, E., Win, M.N., Regitsky, D.D., Santos, C.N.S., Kim, P.B., Cooper, S.R., Raisner, R.M., Herman, A., Sivitz, A.B.: An engineered microbial platform for direct biofuel production from brown macroalgae. Science **335**(6066), 308–313 (2012)
59. Yantovski, E.I.: The solar energy conversion through seaweed photosynthesis and zero emissions power generation. Surf. Eng. Appl. Electrochem. **44**(138), (2008)
60. Zilberman, D., Rajagopal, D., Kaplan, S.: Effect of biofuel on agricultural supply and land use. In: Khanna, M., Zilberman, D. (eds.) Handbook of Biofuel. Springer, New York

Maritime Search and Rescue (MSAR) Operations: An Analysis of Optimal Asset Allocation

Erhan Akbayrak and Mustafa Kemal Tural

Abstract Managing sea rescue assets and their distribution in the various search and rescue (SAR) locations should be carried out according to some well-defined criteria in order to cover SAR areas of the world properly. In this chapter, we intend to give a comparative literature review of maritime search and rescue (MSAR) operations with more emphasis on asset allocation. A framework of a new approach that aims to locate SAR stations and allocate SAR assets to the stations is introduced to the MSAR literature that takes into account the traffic density of sea and air roads over an SAR responsibility area. The model itself is not covered within this book chapter, but instead its main features are discussed and the differences and novelties with respect to the modeling approaches generally used in the literature are presented.

Keywords Search and rescue · SAR · Maritime search and rescue · Maritime rescue · Global maritime distress and safety · Sea rescue

1 Introduction

According to World Health Organization (WHO) media center fact sheet [36], drowning is the 3rd leading (7% of all injury related deaths) cause of unintentional injury death worldwide. It is estimated that approximately 372,000 drowning deaths occur annually. A similar statistic is given by Golden and Tipton [12] which reveals that each year approximately 140,000 water-related deaths happen worldwide. Even though the estimated casualties show the importance of the maritime search and rescue (MSAR) operations, regardless of the numbers given above, a single life is impor-

E. Akbayrak (✉) · M.K. Tural
Department of Industrial Engineering, Middle East Technical University,
06531 Ankara, Turkey
e-mail: erhan.akbayrak@metu.edu.tr

M.K. Tural
e-mail: tural@metu.edu.tr

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_2

tant when planning an MSAR operation. The first thing we will examine about the MSAR is the conventions that regulate rescue at sea.

The international convention regulating the search and rescue (SAR) operations worldwide is “The International Convention on Maritime Search and Rescue (Hamburg, 1979)” [17]. According to the 1979 Convention which entered in force in 1985, it is intended to develop an international SAR plan, so that, no matter where a sea accident happens, the rescue of persons in distress at sea will be coordinated by an SAR organization and, when necessary, by co-operation between neighboring SAR organizations. After the adoption of the 1979 SAR Convention, International Maritime Organization (IMO)’s Maritime Safety Committee segmented the world’s seas/oceans into 13 SAR areas, in each of which the concerned countries have their own SAR responsibility areas. Within these SAR areas, upon a rescue operation is initiated, how well it is managed by the rescue centers depends mainly on the allocation of the rescue assets.

Most of the MSAR operations are initiated by a distress signal received by any means of communication. Sometimes, without having a distress signal, the information of a ship or an aircraft not reaching its final destination within a certain time period can initiate the MSAR operations. High sea conditions, illegal immigration, ship collisions, man overboard etc. are among the major causes of the sea-accidents.

When a man or any floatable object goes overboard at sea, without propulsion, it is subject to drift due to waves, sea currents and wind conditions.[1] Accurately locating the position of such a drifting target without RFID, Global Maritime Distress and Safety System (GMDSS),[2] etc. needs a reliable estimation of the environmental conditions and the target’s last known position. If no reliable last know position is provided when rescue center (RC) is informed about the accident, then, all SAR responsibility areas should be searched. The key questions arising from the uncertainty are “How should the area be searched?”, “Where should the SAR stations be located?” and “How should the SAR assets (helicopters, ships, fixed wing aircrafts, etc.) be allocated to shorten the rescue time?” With this inspiration, we intend to give a literature review of maritime search and rescue (MSAR) operations with more emphasis on asset allocation.

In Sect. 2 of this chapter, we review the literature on the MSAR and discuss their shortfalls. In Sect. 3, we describe the basics of a model for the location of SAR stations and the allocation of SAR assets, without going into the mathematical and modeling details. In Sect. 4, possible extensions to the chapter are discussed, and in Sect. 5, we give some concluding remarks.

[1] The wind speed by itself is a key factor to generate surface waves and currents. Surface currents can be estimated to be approximately 2% of the wind speed when navigating close to the shore. The single affect of the wind over sea currents is itself a research area in the SAR literature [8] and will not be addressed here.

[2] The Global Maritime Distress and Safety System (GMDSS) [11] is the international radio safety system mandated by the International Maritime Organization (IMO) for vessels at sea. The GMDSS was implemented on February 1, 1999 through amendments to the Safety of Life At Sea (SOLAS) Convention. The main aim of GMDSS is to organize and improve emergency communications for the world’s shipping industry.

2 Literature Review

Search and rescue operations can be conducted during either wartime or peacetime. The literature is divided into two by this separation. Although there is a significant amount of research about combat search and rescue [25] in the literature, this is outside the scope of this chapter.

One can categorize the SAR literature into three parts by considering peacetime MSAR operations. In general, these research areas are about,

- Locating SAR stations,
- Leeway (the motion of a drifting object) formulations and search plans[3] to locate the objects under consideration,
- Allocating SAR assets.

Almost all of the studies in the literature on the research areas above deal with a time called “rescue time” and/or with a plan called “search plan”. However, no integrated model encountered in the MSAR literature combining all the three research areas listed above. In fact, minimizing rescue time, locating SAR stations, allocating SAR assets with respect to available/required SAR stations, search plans, budget, areal density of sea and air accidents, etc. are the cornerstone of solving such a complex real-life support problem.

In addition, no MSAR modelling approach in the literature covers both sea accidents and air accidents occurred over seas in the same model. There are separate modeling examples, but not addressing both dimensions. Our proposed modelling approach projects air roads into the sea area and deals with both sea and air accidents over seas simultaneously.

In the MSAR literature, one is more likely to find papers/theses written about locating SAR stations than allocating the SAR assets. One of them is written by Basdemir in 2000 [3] that aims to find the optimal location of new SAR stations. The optimum SAR locations are found by solving a maximal covering location problem. Main emphasis is given to find the minimum number of SAR locations that achieves maximum coverage in the operation area. The SAR locations considered by Basdemir are just for locating SAR helicopters. No other assets other than helicopters are considered in the paper. A similar research is conducted by Haagensen et al. [14] who examined Long-range Rescue Helicopter Missions in the Arctic. A comprehensive model should cover not only SAR helicopters but also other essential rescue assets used in an integrated MSAR operation like fixed wing planes, rescue ships and even unmanned aerial vehicles.

Some part of the literature about MSAR is engaged with the leeway formulations. One of them is written by Oyvind and Allen in 2008 [27] in order to provide a new

[3] A search plan is the method that SAR assets will use in the SAR area. Various types of search plans are introduced to the literature so far, but the one that attracts most of the attention is the CASP (The United States Coast Guard Computer Assisted Search Planning System) [30]. The CASP methodology is mainly based upon Monte Carlo simulation to obtain an initial probability distribution for target location and to update this distribution to account for drift due to currents and winds.

operational, ensemble-based search and rescue model for the Norwegian Sea and the North Sea. A new, robust formulation for the relation between the wind and the motion of a drifting object (leeway) is provided by this paper. Another research about leeway computations is "Optimal Search for a Moving Target: A Geometric Approach" by Snorrason and Ablavsky [31]. Coming up with a path-planner that simultaneously optimizes path efficiency and search-effort allocation is the main contribution of this research. One of the significant properties of the research results is that the paths calculated are also suitable for unmanned aerial vehicles (UAV). The use of UAVs in SAR operations is getting more frequent nowadays, and in our point of view, this fact keeps this particular research up-to-date.

One part of the leeway formulations is to determine an optimal search pattern for a lost target. Optimal search algorithms commonly use Bayesian approach. One of them is "Optimal Search for a Lost Target in a Bayesian World" written by Bourgaut et al. in 2006 [5]. This paper mainly presents a Bayesian approach to the problem of searching for a single lost target by a single autonomous sensor platform. The target may be static or mobile but not evading. The mean time to detection and the cumulative probability of detection are the two significant factors they try to determine by using the Bayesian approach.

It may not be an exaggeration when one claims that one of the greatest contributions to the literature of allocating SAR assets is given by Abi-Zeid and Frost [1]. They present SARPlan, a geographic decision support system designed to assist the Canadian Forces in the optimal planning of search missions for missing aircraft. Its primary purpose is to ensure that the available search resources are deployed in a way that will maximize the mission's probability of success. The optimization modules are based on search theory, gradient search methods, and constraint satisfaction programming. Results demonstrate that SARPlan improves the performance when compared to the manual method. In 2001, SARPlan was the winner of three prestigious excellence awards in the information technology domain. Even though, SARPlan is mainly developed to search missing aircrafts (which is a kind of weakness of the model), it can be applied to find other missing objects of similar size. There are more than a dozen MSAR search methods like expanding square, parallel search, etc. Deciding on the search plan to be used in an MSAR operation is an essential part in the optimization of the rescue time.

Another research about planning sea rescue resources and their distribution in various locations is conducted by Azofra et al. [2]. In the research, they emphasize the difficulties that can be faced when allocating SAR assets in countries with autonomous regional governments. Their main contribution to the literature is realized by formalizing a general methodology based on gravitational models which can be used to define individual and zonal distribution models. They define the following five factors in the process of assigning sea rescue resources:

- Characteristics of the accident, the vessel and the damage produced,
- Types of the accidents and establishment of a scale of severity,
- Distribution of resources such as helicopters, tug-boats and rescue boats with a definition of their radius of action,

- Placement of resources assigning indicators of suitability to locations,
- Cost-effectiveness.

They also introduce the Zonal Distribution Model which considers the zones of the sea whose size means that the conditions of access to any of its geographical position must be quite similar. They name accidents occurring in each of these zones as “superaccident”. With this approach, they categorize the sea area into zonal parts just considering the occurance rate of sea accidents. They also further categorize the accidents with respect to their severity into five category (superaccident, very serious, serious, moderate, and slight). They state that all these five types of accidents happen close to each other within a zone. The likelihood of accidents within the SAR area are based on the historical accident data of the region. Point-wise zonal accident density is the key part within their study.

When we examine the Zonal Distribution Modelling introduced by Azofra et al., we realize that “choke points” of sea accidents do not explain the gravitational modelling needs entirely. The historical data of accident locations does not entirely explain the point estimation of future sea accidents. In our model, we also plan to benefit from the traffic density of sea and air roads over an SAR responsibility area. It can be clearly seen from an instant map presented by www.marinetraffic.com [23] that the traffic density of sea roads does not follow a point-wise zonal pattern, instead, a rectangular pattern compatible with shortest paths is followed by the ships.

Most of the location models on the MSAR are summarized in the thesis by Li [21]. A maximal covering location problem model with weights on incident classes, workload capacities and stochastic considerations are discussed within the thesis. Response times, workload, vessel utilization, and locating rescue vessels are studied with a range of 5–50 rescue vessels.

Another useful reference about the MSAR operations is the Coast Guard Addendum (2013) to the United States National Search and Rescue Supplement [34], which is a supplement to the International Aeronautical and Maritime Search and Rescue Manual. This Addendum establishes policy, guidelines, procedures and general information for Coast Guard use in search and rescue operations. The manual covers the following topics in general:

- Search and Rescue (SAR) organization,
- SAR agreements,
- International SAR,
- General SAR policies,
- SAR communications,
- Rescue planning and operations,
- Coast guard SAR units,
- Procedures for underwater incidents,
- Search planning guide.

It is one of the few references on the procedures for underwater incidents. In addition to that, various types of theoretic SAR data can be found within the manual (i.e., visual search altitudes, height of eye versus horizon range).

There are three famous national search models (U.S., Canadian, and Norwegian) and all the three search models are summarized by Hillier [15]. Inputs for SAR planning and environmental data sets (i.e., types of currents, wind forecasts) are two important topics covered in separate chapters in this work.

Malik et al. in [22] present a joint work with the U.S. Coast Guard's Ninth Disctrict and Atlantic Area Commands where they developed a visual analytics system to analyze historic response operations and assess the potential risks in the maritime environment associated with the hypothetical allocation of coast guard resources.

According to our assessment, evaluation of the risk should be supported by the real time areal usage of the region. Our proposed model will use in addition to the historic data, the traffic density of the sea and air roads for a given region.

A multi-objective model, Incident Based-Boat Allocation Model (IB-BAM), for allocating search and rescue boats is proposed by Razi et al. in [29]. The model consists of three parts. First, by using the Analytic Hierarchy Process (AHP), it determines the weight of each incident type considering its severity. Second, considering historical incident data, a Zonal Distribution Model generates aggregated weighted demand locations. Third, a multi-objective mixed integer program determines locations and responsibility zones of search and rescue boats. Azofra et al. [2] and this particular article both use Zonal Distribution Model and categorize the accidents with respect to their severity. However, neither of them considers the real-time areal density of sea and air roads. They just rely on the historical data of the incidents. In fact, historical data of incidents itself can not explain all the risk associated with future accidents.

3 Model Structure

Our MSAR literature review shows that the following MSAR modelling approaches have been mainly used/developed so far.

- Location modelling of SAR stations [3],
- Allocation modelling of SAR assets [2, 14, 15, 21, 24, 27, 29, 30],
- Risk assessment modelling of SAR areas [4, 13, 22, 35],
- Search theory and SAR planning modelling [1, 5–10, 12, 15, 16, 18–20, 25, 27, 28, 30–33].

We realized the gap in the MSAR literature that there is no integrated or comprehensive modelling approach dealing with locating SAR stations, allocating SAR assets to these SAR stations, making risk assessment of the SAR area, and finally SAR planning to rescue the casualties within an acceptable time frame at the same time. Also, narrowing the scope of our review, to the best of our knowledge, MSAR modelling approaches do not consider airspace over seas as a source of incident. Obviously, air accidents over seas should be taken into account by using historical data of air incidents and the real-time areal density of air roads over seas.

Risk assessment models mainly benefit from past data of sea accidents of a certain SAR region [22]. After mapping the events occurred before, these models forecast a future risk of the SAR area without considering real-time air and sea road traffic densities of the area under consideration.

It is a well known fact that by allocating any kind of asset to any proper station, we intend to maximize the utility of the assets to be assigned. Before stating the intended framework of our model structure, we need to say a few words about "how can SAR assets' utility be assessed?" A short answer would be "The closer to the SAR area (the theatre, i.e., the accident location), the greater the utility of an SAR asset will be." One aim of our model is to distribute SAR assets (ships, helicopters, fixed wing aircrafts, etc.) to SAR stations in such a way that the distance between the SAR stations and the possible accident locations would be the minimum. Since the location of any possible accident is unknown and to be predicted, we intend to focus on traffic density of sea[4] and air[5] roads together with the past data of the accidents occurred before. We will map the real-time traffic density area of sea and air roads [23] and the historical data of accidents occurred on the same chart in order to assess the risk of a particular SAR region. This is a new risk assessment approach to the literature of the MSAR. Most of the models in the literature are generated under the assumption that sea accidents occur in a uniformly manner, or based on the limited historical data. The main concern of most of the models provided so far is to cover all SAR responsibility area within a reasonable time frame. We will not cover explicitly our model structure here, since it is a part of an ongoing research.

In our model, we would like to decide on the number of SAR stations and where to locate them, the allocation of the SAR assets to the SAR stations, the risk assessment of the SAR area and possible search plans to be conducted. We consider the traffic density[6] of each air and sea roads within the SAR area. There can be two kinds of accident possibility related with the sea. First, a vessel could sink due to an accident or rough sea conditions (ship-based accidents). Second, an aircraft flying over the sea could crash or land over sea (aircraft-based accidents). We do not consider swimming-activity-based accidents close to the shores in the model, since these individual accidents do not require time consuming rescue efforts frequently.

The objective of the model is to minimize the average rescue time while having the constraints:

[4] The Sea Roads: These roads are invisible to notice at sea, but followed by most of the sailors, since they provide the shortest distance between departing and last port of calls. Most of the navigation charts does not include them unless you are sailing under a regulated straight or a narrow channel.

[5] The Air Roads: These roads are invisible to notice in the air, but followed by all of the civilian pilots, since there is a multinational convention (Convention on International Civil Aviation by International Commission for Air Navigation (ICAN), 1944, Paris) regulating the air safety. All of the aviation charts include them and, it is mandatory for civilian aircrafts to follow these air roads.

[6] i.e., the traffic density rate of sea and air roads in the ith station SAR responsibility area. (It is the rate of ships and aircraft in the area per hour). Arrival of the ships and aircraft is assumed to be a Poisson process. It is assumed that an aircraft/ship can use either side of their route within a 5 NM buffer zone. These 10 NM-width areas along the routes are assumed to be subject to the highest risk of an air/sea accident. In short, not every part of the SAR area has the same probability of accident occurance.

- Cover all MSAR responsibility area with the search capacity of all assets,[7]
- Decide on the number of SAR stations and allocate each SAR asset to the proper station,[8]
- Consider national regulations on the assets and the stations,
- Consider available budget.
- Consider the risk assessment by using historical data of accidents and the real-time air and sea road traffic densities.

The following assumptions are taken into account:

Assumption 1 In order to determine the percentage of ith station SAR responsibility area subject to the sea and air roads, it is assumed that a ship or an aircraft can use either side of this route within a 5 NM buffer zone. These 10 NM-width buffer areas along the routes are assumed to be subject to the highest risk of sea and air accidents.

This assumption is based on the observation of routes used by the commercial ships and aircrafts. It is observed that a great majority of the ships tend to use shortest path between their departure port and last port of calls. The spread from the center of the route course is assumed 5 NM to the right (starboard) and 5 NM to the left (port) side.

A similar assumption is made for the airlines. But, this time, the center route courses (head of the plane) are regulated by ICAO, i.e., by air maps. These roads may not be the shortest path between departing airport and the destination due to the air safety. The spread from the center of the route course is also assumed 5 NM to the right and 5 NM to the left.

Assumption 2 In order to determine area coverage factor of a single SAR asset, it is assumed that the rescue operation is being conducted under day time and good visibility conditions.

Day time or night rescue operations under good or poor visibility have different probabilities to detect an object at sea. We can further categorize the visibility and sea state on a numeric scale as in the U.S. Coast Guard addendum [34].

Next, we intend to generate a simulation method to test average rescue times with respect to positions of the accidents within the MSAR area. Since the SAR assets are dependent to each other, especially the ship-airborne helos, the area to be reached by

[7] Area coverage factor of a single rescue ship, a helicopter and a fixed wing aircraft: A rescue ship is assumed to have $25kts.$ (nautical miles per hour) permanent cruise speed and a 2 NM-radius detection range. In an hour, a ship will sail 25 NM and will cover a 4 NM-width rectangle area, which equals 100 NM^2.
A rescue helicopter is assumed to have $90kts.$ permanent speed and a 5 NM-radius detection range. In an hour, a helicopter will fly 90 NM and will cover a 10 NM-width rectangle area, which equals 900 NM^2.
A rescue fixed wing aircraft is assumed to have $200kts.$ permanent speed and a 5 NM-radius detection range. In an hour, a fixed wing aircraft will fly 200 NM and will cover a 10 NM-width rectangle area, which equals 2,000 NM^2.

[8] Proportion of SAR responsibility area subject to the air/sea roads are used to differentiate the importance of the SAR areas under concern.

each asset should be simulated in terms of their operation radius considering their fuel replenishment at sea or ashore.

4 Further Discussions

In the preceding section, we stated the framework of a model that takes into account sea and air roads. What if a rescue center in a SAR station does not have suitable geographical formation to accommodate an SAR asset (e.g., high mountains along the shore may be an obstacle for rescue aircrafts to operate.). Then, such assets will not be assigned to this SAR station and they have to operate from a neighboring SAR station. As an extension, one may also consider the negative affects of environmental difficulties. In such cases, the rescue times will increase. In addition to that, it is a common practice for neighboring SAR stations to cooperate in case of a sea accident which is not taken into account in the current version of our model.

On the other hand, if an SAR region has islands or man-made platforms at sea, this will shorten the rescue times for two reasons. First, SAR assets can be located at these places and reaction time will decrease eventually. Second, a platform or a geographical formation (an island, etc.) may act as an SAR asset by itself, since they have an observation range over the sea.

5 Summary and Concluding Remarks

Managing sea rescue resources and their distribution in the various locations should be conducted according to well-defined criteria in order to cover the 13 SAR areas of the world. Decision makers should be urged to optimize their MSAR efforts by benefitting from various allocation algorithms.

The reason for us to write this chapter is to provide a brief literature review and give a new approach of allocating SAR assets. As of our knowledge, especially the air road density over seas is not considered by any other paper or book written before. According to The National Air Traffic Controllers Association (NATCA) of USA [26], at any given moment, roughly 5,000 planes are in the skies above the United States (approximately 70,000 flights daily). Looking at these numbers, we can evaluate the importance of the air roads when considering air-based accidents over seas.

A possible extension of the model stated can be to predict the traffic intensity rates of sea and air roads. Here, we use marinetraffic.com real time merchant ship data which excludes navy ships [23]. Because of this missing data, the actual traffic densities of sea roads may be higher than the ones we used in the model.

The results found are expected to change according to the visibility and day & night conditions of the SAR areas. Poor visibility or a night SAR operation will result in an increased rescue time and more rescue assets.

References

1. Abi-Zeid, I., Frost, J.R.: SARPlan: a decision support system for canadian search and rescue operations. Eur. J. Oper. Res. **162**(3), 630–653 (2005)
2. Azofra, M., Peres-Labajos, C.A., Blanco, B., Achutegui, J.J.: Optimum placement of sea rescue resources. Saf. Sci. **45**(9), 941–951 (2007)
3. Basdemir, M.M.: Locating Search and Rescue Stations in the Aegean and Western Mediterranean Regions of Turkey. Air Force Institute of Technology, Wright-Patterson AFB OH (2000)
4. Bichler-Robertson G.M.: Maritime commercial passenger ship casualties, 1950 -1998 an analysis of negligent corporate risk-taking and system hazard, Ph.D. thesis, Newark, Rutgers - The State University of New Jersey (2000)
5. Bourgaut, F., Furukawa, T., Durrant-Wayte, H.F.: Optimal search for a lost target in a Bayesian world. Field Serv. Robot. **24**, 209–222 (2006)
6. Breivik, O., Allen, A.: An operational search and rescue model for the norwegian sea and the north sea. J. Mar. Syst. **69**(1–2), 99–113 (2008)
7. Breivik, O., Allen, A., Maisondieu, C., Olagnon, M.: Advances in search and rescue at sea. Ocean Dyn. **63**(1), 83–88 (2012)
8. Burciu, Z.: The influence of wind speed on surface water currents and its reference to search and rescue actions at sea. Arch. Transp. Pol. Acad. Sci. Comm. Transp. **14**(2), 39–50 (2002)
9. Chapline W.E.: Estimating the drift of distressed small craft, coast guard alumni, association bulletin, U.S. Coast Guard Academy, New London, CT, **22**(2), pp. 39–42 (1960)
10. Frost J.R., Stone L.D: Review of search theory: advances and applications to search and rescue decision support. Technical report CG-D-15-01, US Coast Guard Research and Development Center, Groton, CT, USA (2001)
11. Global Maritime Distress and Safety System. www.gmdss.com (2014). Accessed 30 Dec 2014
12. Golden, F., Tipton, M.: Essentials of Sea Survival. Human Kinetics, Leeds, 1st edn. (2002)
13. Guedes Soares, C., Teixeira, A.P.: Risk assessment in maritime transportation. Reliab. Eng. Syst. Saf. **74**(3), 299–309 (2001)
14. Haagensen, R., Sjoborg, K., Rossing, A., Ingilae, H., Markengbakken, L., Steen, P.A.: Long-range rescue helicopter missions in the arctic. Prehospital Disaster Med. **19**(2), 158–163 (2004)
15. Hillier, L.E.: Validating and improving the Canadian coast guard search and rescue planning program (CANSARP) ocean drift theory, Department of Environmental Science, Memorial University of Newfoundland (2008)
16. Hufford, G.L., Broida, S.: Estimation of the leeway drift of small craft. Ocean Eng. **3**(3), 123–132 (1976)
17. International Convention on Maritime Search and Rescue, International Maritime Organization (IMO), 1979, Hamburg. www.imo.org (2014). Accessed 30 Dec 2014
18. Koopman, B.O.: The theory of search, part II: target detection. Oper. Res. **4**(5), 503–531 (1956)
19. Koopman, B.O.: The theory of search, part III: the optimum distribution of searching effort. Oper. Res. **5**(5), 613–626 (1957)
20. Koopman, B.O.: Search and Screening: General Principles With Historical Applications. Pergamon Press, New York (1980)
21. Li L., Rescue Vessel Location Modelling, M.A.Sc., Department of Industrial Engineering, Dalhousie University, Canada (2006)
22. Malik, A., Maciejewski, R., Maule, B., Ebert, D.S.: A visual analytics process for maritime resource allocation and risk assessment. In: IEEE Symposium on Visual Analytics, Science and Technology (2011)
23. Marine Traffic. www.marinetraffic.com (2014). Accessed 30 Dec 2014
24. Marven C.A., Canessa R.R., Keller P.: Exploratory spatial data analysis to support maritime search and rescue planning. In: Geomatics Solutions for Disaster Management, pp. 271–288. Springer, Berlin (2007)
25. Moentmann J., Holland E., Wolver G.: Joint combat search and rescue-operational necessity or afterthought? Jt. Forces Q., 44–51 (1998)

26. National Air Traffic Controllers Association (NATCA): USA. http://www.natca.org (2014). Accessed 30 Dec 2014
27. Oyvind, B., Allen, A.A.: An operational search and rescue model for the Norwegian Sea and the North Sea. J. Mar. Syst. **69**(1–2), 99–113 (2008)
28. Pingree, F.: Forethoughts on rubber rafts. Technical report, Woods Hole Oceanographic Institution (1944)
29. Razi, N., Karatas, M.: A multi-objective model for locating search and rescue boats. Eur. J. Oper. Res. **254**(1), 279–293 (2016)
30. Richardson, H.R., Discenza, J.H.: The United States coast guard computer-assisted search planning system (CASP). Nav. Res. Logist. Q. **27**(4), 659–680 (1980)
31. Snorrason M., Ablavsky V.: Optimal Search For a Moving Target: A Geometric Approach, American Institute of Aeronoutics & Astronoutics (2000)
32. Stone, L.D.: Theory of Optimal Search, 2nd edn. Operations Research Society of America, Arlington (1989)
33. Suzuki, T., Sato, H.: Measurement of the drifting of a fishing boat or research vessel due to wind and wave. J. Jpn. Inst. Navig. **65**(4), 1225–1245 (1977)
34. U.S. Coast Guard: U.S. Coast Guard addendum to the United States national search and rescue supplement to the international aeronautical maritime search and rescue manual, United States Coast Guard Publication (2013)
35. Wang, J.: The current status and future aspects in formal ship safety assessment. Saf. Sci. **38**(1), 19–30 (2001)
36. World Health Organization: Drowning Fact Sheet N347, updated November 2014. http://www.who.int/mediacentre/factsheets/fs347/en/ (2014). Accessed 30 Dec 2014

Measuring the Regional Productivity in Spanish Hotel Industry

P. Alberca and L. Parte

Abstract This paper analyzes regional productivity in the Spanish hotel industry and the influence of regional and management factors on productivity. The study develops a novel approach to hotel productivity by attempting to establish a relationship with regional and management performance variables. In terms of methodology, this study first applies the nonparametric Malmquist technique to microeconomic balanced panel data in order to measure Total Factor Productivity (*TFP*). Subsequently, two groups of key factors are introduced as possible determinants of the productivity change: regional industry factors and hotel company management factors. The results suggest differences in regional productivity across the sample.

Keywords Performance evaluation · Total factor productivity · Regressions analysis · Panel data · Hotel firms · Decision making process

Highlights

- This paper analyzes Total Factor Productivity (*TFP*) of Spanish hotel firms.
- The study introduces specific variables to investigate differences in productivity.
- The study examines how regional and management factors influence *TFP*.
- The Malmquist nonparametric technique is used to measure productivity change.
- OLS regression is used to test the relationship between TFP and both regional and management factors.

P. Alberca (✉) · L. Parte
Department of Finance and Accounting, Faculty of Economics,
Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
e-mail: palberca@cee.uned.es

L. Parte
e-mail: lparte@cee.uned.es

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_3

1 Introduction

Over the past six decades, tourism has undergone continuous expansion and diversification, becoming one of the largest and fastest-growing economic sectors in the world [39]. Despite occasional shocks to the sector, international tourist arrivals have recorded virtually uninterrupted growth. In 2014, tourism was the leading industry in the Spanish economy, contributing 11.4% of Spanish national Gross Domestic Product (*GDP*).

A matter of considerable interest in the tourism industry is firm efficiency and productivity. Therefore, developing tools in order to evaluate the efficiency and productivity of the tourism activities is an important issue. Moreover, understanding the influence of regional and managerial factors on productivity can help managers to define their business strategy in the hotel industry. Policymakers are also interested in promoting the positive effects of competition and providing guidance on how to correct inefficient management directions [6].

Methodologically, productivity measurement can be approached in different ways. Within the tourism industry the literature has mainly focused on efficiency analysis using both parametric and nonparametric frontier methodologies [3, 7, 10, 19, 22].

However, a criticism of these efficiency studies relates to the fact that the empirical results may not consider for both technological and efficiency change over time, i.e. they do not answer whether efficiency improves or deteriorates over time because efficiency is only measured with respect to a specific year. To address this concern, some studies use alternative models to measure productivity change, mainly nonparametric methodologies such as the Malmquist Total Factor Productivity (*TFP*) approach [2, 9, 17, 23, 28, 43], the Stochastic frontier approach [11, 27] and Luenberger productivity indicator [18].

In general, the literature suggests that productivity in the hotel industry is low in comparison to other industries [13, 24, 27, 37]. Also, findings show that the productivity index decreases, which is almost always explained by technological change deterioration (e.g., [18] using a sample of French ski resorts).

This study examines the hotel productivity change in all regions of Spain and substantially differs from the above mentioned literature in several aspects. To date, one of the primary shortcomings of recently published microeconomic studies is related to the use of a low number of observations. Other efficiency studies use aggregate dates for a country, region, or other geographical entity and most Data Envelopment Analysis (DEA) in the hospitality industry focus on hotel efficiency at the cross-sectional level. In order to overcome this limitation, we consider microeconomic panel data disaggregated by region.

Specifically, the paper uses disaggregated data on a large sample of Spanish hotel firms over a long period of time to examine both the Total Factor Productivity (*TFP*) index and regional differences. The sample is composed of balanced panel data for 1,593 hotel firms and covers the period from 2001 to 2008. As the current financial crisis started in 2007, in this study we consider a possible break in the productivity tendency.

In terms of methodology, the study applies the Malmquist index of TFP using distances measures relative to DEA frontiers [14]. DEA frontier methodology is used in the main area of research on frontier modeling alongside the stochastic frontier models, and both models are useful in analyzing efficiency. However, DEA is preferred when the researcher has doubts as to which functional form to adopt [5]. The Malmquist index of TFP and the decomposition proposed by Färe et al. [14] is used as the analytical tool in order to assess productivity change and its main components: technical change and efficiency change.

Previous literature focusing on productivity determinants shows that quality in labor and capital inputs, information technology and research and development, spillovers, competition, deregulation or proper regulation are key influencing factors of TFP. Syverson [36] provides an excellent review of this subject. The author considers there to be two groups of variables that explain TFP: business or internal factors, which come under the control of managers; and environmental or external factors, which are more difficult for managers to control.

Based on the findings of this literature, we explore the association between productivity change and both business factors and regional industry factors. Accordingly, the second objective of this paper is to investigate the influence of a number of different factors on the performance and productivity of hotel firms across all Spanish regions. We use OLS regression to test the association and productivity of hotel firms across all Spanish regions.

This paper is organized as follows. Section 2 presents a brief review of the literature. Section 3 provides the research design of the study, specifically, the formal hypotheses, the empirical models and the determinants of TFP. Section 4 describes the sample and provides the descriptive statistics, Sect. 5 presents the results of the empirical models and, finally, Sect. 6 discusses the results and presents the overall conclusions.

2 Literature Review

Prior literature has used a wide variety of empirical models to evaluate firm performance. Specifically, DEA methodology has used to assess efficiency and productivity along countries and industries as tourism [1, 20–22], commercial banks [25] or European banks [17], among others.

Within the tourism industry, previous studies have mainly focused on efficiency analysis using both parametric and nonparametric frontier methodologies [1, 3, 6, 7, 10, 19, 20, 22]. Furthermore, many studies have concentrated on what activities generate most efficiency. For example, [22] examine hotel activities or business units that generate the highest efficiency. They use a two-process (or two-stage) framework to calculate the efficiency driven by occupancy activity and catering activity. This division allows a better measure of value-added chain and resource allocation in hotel firms. Hsieh and Lin [20] use relational network data envelopment to analyze the efficiency and effectiveness of international tourist hotels in Taiwan.

However most empirical studies only measure the efficiency in one specific year, and then, the analysts may not ascertain whether efficiency improves or deteriorates over time [21]. In order to overcome this problem, some studies have focused on measuring the TFP in the tourism industry [2, 4, 11, 13, 18, 23, 24, 27, 28, 37].

Generally, prior studies that examine productivity change use the Malmquist productivity index based on nonparametric data envelopment analysis (DEA). For example, Fiordelisi and Molyneux [17] in the banking industry; Chang et al. [9] in US firms in order to examine productivity before and after the Sarbanes–Oxley Act; and Barros and Alves [2], Barros [4], Hwang and Chang [23] and Yu and Chen [43] in the hotel industry. Yu and Chen [43] consider that the use of metafrontier Malmquist productivity index has several advantages. In addition to satisfying the requirement of circularity and being immune to linear programming infeasibility, the index also considers the heterogeneity among hotels and overcomes the problem of base period dependency.

In this research, the Malmquist's nonparametric technique is applied to microeconomic panel data in order to measure productivity. The Malmquist productivity change index assesses the specific position of a firm and also measures the change in productivity. Our research also extends prior studies focused on the influence of particular variables in terms of improving efficiency and productivity. Therefore, the paper introduces two groups of key factors as possible determinants of the productivity change, namely regional industry factors (tourist flows) and firm management factors.

The firm size and location have drawn the attention of researchers and academics in hotel industry [3, 17, 23]. Findings suggest that the hotel size [3, 10, 23, 30, 32] and the location [3, 31, 33, 41] are associated with the efficiency score.

Occupancy rates [12, 22, 31, 35, 41], growth rates in total foreign tourist arrivals [12, 22, 23, 31], macro socioeconomic factors [21] and factors such as tourism attractions, business environment, tourist shopping attractions, information and transportation conditions, level of tourism professionalism, degree of information and promotion, local amenities, and transport accessibility [41] are also linked to superior hotel efficiency and performance.

Recently Luo et al. [28] showed that variables such as political hierarchy, degree of openness, and level of tourism dependence explain the cross-city differences in efficiency scores while ownership structure is the factor that contributes to boosting efficiency over time. The authors use the DEA model, Malmquist productivity index and the panel Tobit and linear regression models to obtain the evidence. Yin, Tsai, Wu, [42] also suggest analyzing the hotel life cycle (initial, growth, maturity, and recession and regeneration phases) to gain insight into hotel efficiency. Interestingly, Xu, Wei and Zhao [40] use the three-stage DEA model to examine the influence of social media on operational efficiency. The evidence does not show a significant association between operational performance and social media.

Based on this literature, the current study examines the influence of firm factors and regional factors on productivity using microeconomic disaggregated data of hotel firms. In the next section, we explain the research design.

3 Research Design

We implement several steps to achieve the objectives of the paper. Firstly, the Malmquist nonparametric technique is used to measure productivity change between two periods for a particular firm. It is calculated by means of DEA. This approach makes it possible to compute TFP change and decompose these measures into various components: technical regress or advance (a shift in the production frontier) and technical efficiency change (a movement towards or away from the production frontier).

The second objective is to explain regional differences in the productivity index. To do that, in this study we examine regional tourism factors and firm-specific factors that could be related to productivity. Regional tourism factors include the change in estimated hotel capacity by region (*PLAC*), the change in hotel occupancy rates (*OCCUP*) and the change in GDP. Firm specific factors include the change in leverage (*LEV*) and the change in company size (*SIZE*). The association between productivity change and the possible key factors is measured by multivariate analysis (panel regression).

3.1 Study 1. Productivity Change: Methodology and Variables

The first objective of this paper is to analyze productivity change in the Spanish hotel industry. Relatively few studies have focused on productivity change in the hospitality industry and several authors have expressed concerns about this situation [13, 24, 27, 37]. By segmenting the sample into regions, we explore productivity change at the regional level over the period 2001–2008. As we examine a long time period, three distinct periods are considered in this study: the period of expansion from 2001 to 2006, the period of recession from 2007 to 2008 (that is, the productivity index could vary in these years due to the fact that 2007 is generally thought of as the first year when the financial crisis influences on firm performance), and the period as a whole (2001–2008).

3.1.1 The Malmquist Total Factor Productivity Index

The Malmquist TFP index permits the separation of the ‘catching up’ effect, that is, changes in technical efficiency over time, from ‘technological change’, namely shifts in the best practice frontier over time due to technological progress.

In order to define the Malmquist productivity index, it is necessary to consider the distance function with respect to two different time periods. The distance function is the reciprocal of Farrell’s [16] measure of output technical efficiency, which calculates how far an observation is from the technological frontier, and is defined by Shephard [34] and Färe [14] as follows:

$$D_o^t(x_t, y_t) = \inf\left[\theta^t : (x_t, {}^{y_t}\!/_{\theta^t}) \in L(x_t)\right] = \left(\sup\left\{\theta^t : (x_t, \theta^t y_t) \in L(x_t)\right\}\right)^{-1} \tag{1}$$

$L(x_t)$ represents the production technology for each reference period and $(x_t,\ y_t)$ denote a vector of inputs and outputs, respectively.

The Malmquist Index, introduced initially by Caves, Chritensen and Diewert [8], is an appropriate method to calculate variations in TFP between two time periods, t and $t + n$, and was defined as a ratio of distance function measures. The Malmquist index has several advantages that have contributed to its popularity. First, the Malmquist TFP index is more generic than the Törnqvist index, since it allows for inefficient performance, does not presume an underlying functional form for technology, and does not require data on prices or factor shares. Second, the Malmquist TFP index technique makes it possible to separate efficiency change or the catching up effect (changes in the technical efficiency of each firm or Decision Making Unit–DMU-over time with respect to the best practice frontier), from technical change (shifts in the best practice frontier over time due to technological progress). The reason for decomposing productivity change into components is to identify the sources of productivity growth and use this information to tourism policy.

The Malmquist TFP index evaluates TFP change between two data points by calculating the ratio of the distances of each data point or DMU relative to a common technology. If the period t technology is used as the reference technology, the Malmquist TFP index, between period t, or the base period, and period $t + n$, can be expressed as:

$$M^t(y_{t+n}, x_{t+n}, y_t, x_t) = \frac{D^t\,(x_{t+n}, y_{t+n})}{D^t(x_t, y_t)} \tag{2}$$

Commonly, if the period $t + n$ reference technology is used, the Malmquist index can be defined as:

$$M^{t+n}(y_{t+n}, x_{t+n}, y_t, x_t) = \frac{D^{t+n}\,(x_{t+n}, y_{t+n})}{D^{t+n}(x_t, y_t)} \tag{3}$$

represents the distance from the period t observation to the period $t + n$ technology. A value M greater than 1 indicates positive TFP growth from period t to period $t + n$.

As noted by Färe, Grosskopf and Roos [15], the two index represented in the equations above are only equivalent if the technology is Hicks neutral. So as to avoid the necessity of either imposing this restriction or arbitrarily choosing one of the two technologies (t or $t + n$), the Malmquist TFP index is often defined as the geometric mean:

$$M'(y_{t+n}, x_{t+n}, y_t, x_t) = \left[\frac{D^t(x_{t+n}, y_{t+n})}{D^t(x_t, y_t)} \times \frac{D^{t+n}(x_{t+n}, y_{t+n})}{D^{t+n}(x_t, y_t)}\right]^{1/2} \tag{4}$$

The distance functions in this productivity index can be rearranged to show that it is equivalent to the product of a efficiency change (EFFCH) and technical change (TCH):

$$M'(y_{t+n}, x_{t+n}, y_t, x_t) = \overbrace{\frac{D^{t+n}(x_{t+n},\ y_{t+n})}{D^t(x_t,\ y_t)}}^{Efficiency\,change(EFFCH)} \times \left[\overbrace{\frac{D^t(x_{t+n},\ y_{t+n})}{D^{t+n}(x_{t+n},\ y_{t+m})} \times \frac{D^t(x_t,\ y_t)}{D^{t+n}(x_t,\ y_t)}}^{Technical\,change(TCH)} \right]^{1/2} \tag{5}$$

Färe et al. [14] consider that productivity change can be decomposed into changes in efficiency and technical change, unlike Caves et al. [8], who considered the only source of productivity change to be technical change.

This breakdown is useful because it provides information on the causes of overall productivity change. The first component of the Malmquist index (Eq. 5) is referred to as efficiency change (*EFFCH*) and measures the change in distance between the DMU and the production frontier. A value greater than 1 indicates an increase in efficiency relative to the frontier, while a value less than 1 indicates a decline in efficiency. The second component, technical change (*TCH*), is due to a variation in the production frontier between two periods and thus reflects technological improvement or deterioration.

3.1.2 DEA Variables

Implementing the DEA linear programming method requires selecting inputs (resources or costs) and outputs (goods or services). This selection is a critical issue for research into efficiency. Selection of inputs and outputs should be based on the theoretical production process, although the empirical model may be a simple abstraction of the theoretical production function.

Output is relatively easy to quantify, for example, as room nights or sales revenue. In fact, sales is a variable that is commonly used as an output because it represents a measure of a firm's achievement of its goals [1–3, 19]. Input resources for tourist hotel management include operating costs, personnel, capital and equipment [4, 6]. In accordance with previous studies on the hotel industry, the DEA model is applied using the following variables: sales (as output) and the number of full-time employees to represent manpower, the book value of a property to represent capital investment of hotels and cost of sales to represent the cost of inputs (as input variables). Therefore, the inputs are defined according to a generalized Cobb-Douglas function. The input and output variables are measured in monetary terms, and thus the study consider the price (inflation) effect on sales revenue and operating costs.

3.2 Study 2. The Relationship Between Total Factor Productivity and Regional and Firm Factors

This paper presents an empirical approach to productivity determinants using a sample of 1,593 hotel firms for the period from 2001 to 2008 (balanced panel). We run OLS regressions to examine the relationship between productivity and a set of explanatory variables.

The simplest way to run the regression is using pooled OLS regression. The model is expressed as follows:

$$y_{it} = \beta_1 X_{it} + e_{it} \tag{6}$$

where

y_{it} is the dependent variable, in our case the TFP growth for firm i in year t
X_{it} is a vector of TFP determinants
e_{it} is the error term.

If the model satisfies the conditions of the classical model (zero conditional mean of e_{it}, homoscedasticity, independence across observations, i, and strict exogeneity of x_{it}), then it can be used.

The fixed effects model assumes that the omitted effects in the general model are correlated with the included variables. Firm fixed effects are included to control for unobserved heterogeneity. The model is expressed as follows:

$$y_{it} = \beta_1 X_{it} + a_i + u_{it} \tag{7}$$

where

y_{it} is the dependent variable, in our case the TFP growth for firm i in year t
X_{it} is a vector of TFP determinants
a_i $(i = 1, \dots, n)$ is the unknown intercept for each firm (n firm-specific intercepts)
β_1 is the coefficient for the independent variables
u_{it} is the error term.

Wooldridge [38, 490] explains that "under a strict exogeneity assumption on the explanatory variables, the fixed effects estimator is unbiased: roughly, the idiosyncratic error u_{it} should be uncorrelated with each explanatory variable across all time periods. The fixed effects estimator allows for arbitrary correlation between a_i and the explanatory variables in any time period, just as with first differencing".

The fixed effects models considers time-independent effects for each firm. That is, if the error terms are correlated, then fixed effects is not suitable since inferences may not be correct.

Wooldridge [38, 501] explains that the fixed effects estimator allows for arbitrary correlation between a_i and the explanatory variables, whereas random effects does not, which explains why fixed effects is widely thought to be a more convincing tool for estimating ceteris paribus effects. However, when the key explanatory variable is constant over time, fixed effects cannot be used to estimate its effect on the dependent variable.

Random effects assume that the firm error term is not correlated with the predictors, which allows for time-invariant variables to act as explanatory variables. The specification model is as follows:

$$y_{it} = \beta_1 X_{it} + a_{it} + u_{it} \tag{8}$$

where a_{it} is between-entity error and u_{it} within-entity error.

Many studies use the Hausman test to decide between fixed or random effects models. The idea is to test whether the unique errors (u_i) are correlated with the regressors, with the null hypothesis being that they are not correlated. So, if the Hausman test shows that Prob $> \chi^2$ is less than 0.05, the researcher can use a fixed effects model.

However, as Wooldridge [38, 502] explains, "in practice, a failure to reject means either that the random effects and fixed effects estimates are sufficiently close so that it does not matter which is used, or the sampling variation is so large in the fixed effects estimates that one cannot conclude practically significant differences are statistically significant". In this paper, the Hausman test yields a value of 0.05. We estimate the regression between TFP and explanatory variables using fixed effects.

3.2.1 Regional Variables and Firm Specific Factors

Most previous research in efficiency hotel firms primarily uses cross-section data instead of panel data. Furthermore, recent empirical studies focus on efficiency and productivity determinants. Syverson [36] provides an excellent review of the literature focusing on the productivity determinants. He explains that productivity depends on business factors and environmental factors. The former include drivers such as the managerial practice/talent, higher-quality labor and capital inputs, information technology and research and development, experience, product innovation and firm structure decisions, all of which are under the control of managers. The second set of productivity factors, called environmental determinants, includes drivers such as productivity spillovers, competition, deregulation or proper regulation and flexible input markets. These factors are more difficult for managers to control.

The second objective of this paper is to analyze a set of factors that may be related to hotel productivity change. By breaking down the sample data set into Spanish regions, this study investigates the determinants of hotel productivity at both regional and firm level. The hotel industry is cyclical and consequently is sensitive to shifts in macroeconomics and regional factors. Based on this argument, we consider that changes in economic conditions and especially in tourist flows of a particular region could influence hotel productivity. Accordingly, the productivity of hotel firms may be associated with the tourist flows in the region [12, 41].

Focusing on a specific industry, we include two tourist flows as determinant factors of productivity change: change in estimated Regional Hotel Capacity (*PLAC*) and

Table 1 Definitions of key factors

Regional variables	
Label	Description
$PLAC_k$	Places estimated in the region K in which each company is located
$OCCUP_k$	Hotel occupancy rate in the region K in which each company is located
GDP_k	Gross domestic product by region
Firm specific factors	
Label	Description
LEV_{it}	Leverage of the firm i in the year t
	LEV = Total Debts/Assets
$SIZE_{it}$	Total Asset of the firm i in the year t

change in the Regional Hotel Occupancy Rate (*OCCUP*). We also include change in GDP as a proxy for the total production and services of economy in each region.

Managerial and corporate governance factors define the ways in which the supplier of finance to corporations is assured of getting a return on investment in a firm [29]. Different stakeholders, such as debt holders, equity holders and investors act as pressure groups influencing management strategies. We include capital structure (*LEV*) and company size (*SIZE*) as management factors. The variables definition are shown in Table 1.

4 Sample and Data Descriptions

The financial data were collected from the SABI (*Sistema de Análisis de Balances Ibéricos*) database and information regarding tourist flows was collected from the "Hotel Occupancy Survey". The sample includes data on 1,593 hotel firms for the period from 2001 to 2008. As the sample period covers two economic cycles, it was split into two sub-periods: the 2001–2006 expansion and the 2007–2008 recession.

Table 2 presents the descriptive statistics for the inputs and outputs that were used to evaluate productivity change for the 2001–2008 period. The average output value is increasing and no inflection is exhibited for 2007. The evidence suggests adequate tourist flow behavior in Spain, which also remains high-ranking in terms of world tourism revenue. Additionally, the small size of Spanish firms is notable (for example, the variable Labor input/number of full-time workers). Similar to many other countries, the Spanish hotel industry is composed of both large and small units that compete in the market.

Table 2 also provides descriptive statistics for the firm variables used in the second methodological stage. Table 2 shows that average leverage was around 0.52% and *LEV* decreased over the period. Table 3 reveals the regional tourist flows referred to 2001–2008 period and also for the sub-periods of 2001–2006 and 2006–2008. Inter-

Table 2 Descriptive statistics

Variables	Units	Mean	SD	Maximum	Minimum
2001					
Outputs					
Sales	(thousand €)	2,318.88	6,870.42	222,180.00	48.00
Inputs					
Book value of property	(thousand €)	3,454.80	11,555.87	288,986.00	2.00
Number of full-time workers	(number)	37.94	121.29	4,024.00	2.00
Cost of sales	(thousand €)	515.51	1,443.65	43,731.00	2.00
Balance sheet variables					
Leverage	(%)	0.551	0.559	1.230	0.032
Total assets	(thousand €)	4,066	1,394	344,701.00	3
2006					
Outputs					
Sales	(thousand €)	2,944.34	9,770.45	285,924.00	24.00
Inputs					
Book value of property	(thousand €)	5,159.87	20,371.60	355,482.00	3.00
Number of full-time workers	(number)	44.64	144.87	4,512.00	2.00
Cost of sales	(thousand €)	655.31	1,848.49	45,955.00	2.00
Balance sheet variables					
Leverage	(%)	0.527	0.522	1.120	0.024
Total assets	(thousand €)	5,454	1,824	417,454.70	24
2008					
Outputs					
Sales	(thousand €)	3,062.65	9,923.16	254,778.00	24.00
Inputs					
Book value of property	(thousand €)	6,417.44	38,223.66	1,231,766.00	5.00
Number of full-time workers	(number)	45.25	145.79	4,348.00	2.00
Cost of sales	(thousand €)	645.45	1,843.54	44,078.00	4.00
Balance sheet variables					
LEV leverage	(%)	0.512	0.513	1.028	0.021
Total assets	(thousand €)	7,878	40,382	1,245,733	18

Table 3 Regional tourist flow for the 2001–2008 period

	Change in places estimated (percentage change)			Change in occupancy rate (percentage change)			Change in gross domestic Product (percentage change)		
	2008–2001	2006–2001	2008–2006	2008–2001	2006–2001	2008–2006	2008–2001	2006–2001	2008–2006
Andalusia	0.405	0.313	0.070	−0.057	−0.028	−0.029	0.644	0.500	0.096
Aragón	0.300	0.103	0.179	0.029	0.027	0.002	0.625	0.445	0.125
Asturias/Cantabria	0.290	0.266	0.019	−0.025	0.013	−0.038	0.605	0.439	0.116
Balearic Islands	−0.037	−0.039	0.001	−0.042	−0.009	−0.033	0.556	0.401	0.111
Canary Islands	0.376	0.285	0.071	−0.036	−0.021	−0.016	0.551	0.419	0.092
Castilla-León	0.307	0.200	0.090	−0.023	−0.012	−0.011	0.564	0.417	0.104
Castilla-Mancha	0.251	0.134	0.103	0.001	0.017	−0.017	0.604	0.441	0.113
Catalonia	0.219	0.164	0.047	−0.037	−0.011	−0.025	0.576	0.433	0.100
Valencia	0.407	0.317	0.069	−0.099	−0.049	−0.049	0.600	0.448	0.105
Extremadura	0.170	0.125	0.041	−0.003	−0.002	−0.001	0.699	0.521	0.117
Galicia	0.284	0.210	0.061	0.012	0.032	−0.020	0.592	0.445	0.102
Madrid (Region)	0.474	0.355	0.088	−0.028	−0.001	−0.027	0.610	0.439	0.119
Murcia (Region)	0.308	0.131	0.156	−0.040	−0.012	−0.028	0.609	0.441	0.117
Navarra/La Rioja	0.308	0.205	0.085	−0.025	0.017	−0.043	0.582	0.421	0.113
Basque country	0.383	0.287	0.075	0.012	0.058	−0.046	0.596	0.427	0.118
Mean	0.296	0.204	0.077	−0.024	0.001	−0.025	0.601	0.442	0.110

All variables are defined in Table 1

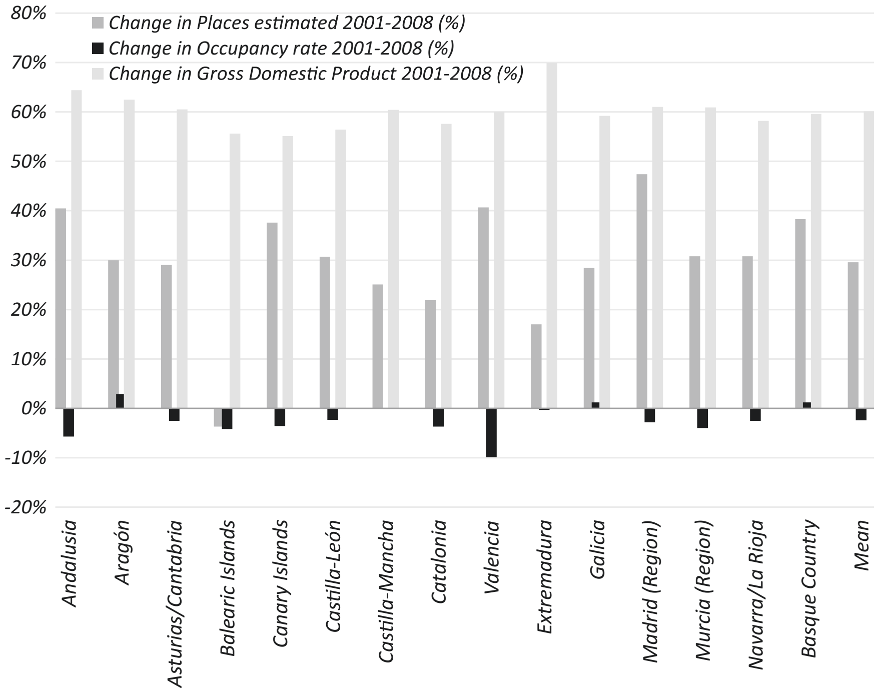


Fig. 1 Change in regional tourist flow 2001–2008 (%)

estingly, *PLAC* and *GDP* increased over the period whereas *OCCUP* decreased in most regions. Figure 1 presents the values graphically referred to 2001–2008 period.

5 Results

5.1 Results of the Study 1

Table 4 summarizes the geometric means of the Malmquist TFP index for all Spanish regions and the full sample of Spanish hotel firms in the period. The first component of the Malmquist index is referred to as efficiency change (*EFFCH*) and measures the change in the distance between a given DMU and the production frontier. The results of the efficiency analysis reveal that (*EFFCH*) increased slightly over our sample period. However, this trend in mean efficiency values cannot be used to deduce total productivity change, because they do not take into account shifts in the efficiency frontier. The second component, technical change (*TCH*), is due to changes in the production frontier between two periods, that is, shifts in company best practices reflect deterioration. The results reveal a decrease in the Total Factor Productivity (*TFP*) in the sample period (−4.2%), which is almost entirely attributable to technical

Table 4 Malmquist index decomposition: summary

Period	*EFFCH*$^{(a)}$	*TCH*$^{(b)}$	*TFP*$^{(c)}$
2001–2006	1.079	0.877	0.946
2006–2008	1.043	0.930	0.970
2001–2008	1.061	0.903	0.958

(*a*) EFFCH Efficiency change
(*b*) TCH Technical change
(*c*) TFP: Total Factor Productivity

Table 5 Malmquist index decomposition. Regional summary (2001–2008)

Region designation	Total DMU$^{(*)}$	Malmquist EFFCH$^{(a)}$	Index TCH$^{(b)}$	Decomposition TFP$^{(c)}$
Andalusia	174	1.0570	0.9025	0.9539
Aragón	63	1.1375	0.9094	1.0344
Balearic Islands	238	1.0744	0.8913	0.9576
Canary Islands	87	1.0039	0.8889	0.8923
Cantabria/Asturias	60	1.0775	0.8982	0.9678
Castilla-León	94	1.0774	0.9093	0.9797
Castilla-Mancha	49	1.0148	0.9098	0.9232
Catalonia	377	1.0731	0.9050	0.9712
Madrid	87	1.0538	0.9149	0.9641
Murcia	16	1.0246	0.9031	0.9252
Valencia	146	1.0521	0.9101	0.9575
Extremadura	24	1.0599	0.9068	0.9609
Navarra/Rioja	40	1.0477	0.8929	0.9355
Galicia	111	1.0141	0.9090	0.9219
Basque Country	27	1.1217	0.9124	1.0233
Geometric means		1.0610	0.9032	0.9582

(*) DMU: Decision making units
(*a*) EFFCH Efficiency change
(*b*) TCH Technical change
(*c*) TFP: Total Factor Productivity-Units: variation rate

change (−9.7%). This finding is consistent with the results obtained by Elfring [13], Johns and Wheeler [24], Witt and Witt [27], Kim [37] and Goncalves [18], among others. These authors conclude that productivity in the tourism industry is comparatively low.

Table 5 shows the Malmquist index decomposition for hotels of the distinct Spanish regions in the sample period and Fig. 2 reports the results of the productivity ranking for the Spanish regions, which can be clearly differentiated into two groups.

The first group includes those regions with productivity growth (only Aragon and the Basque country, with geometric means of 1.03 and 1.02 respectively), while the

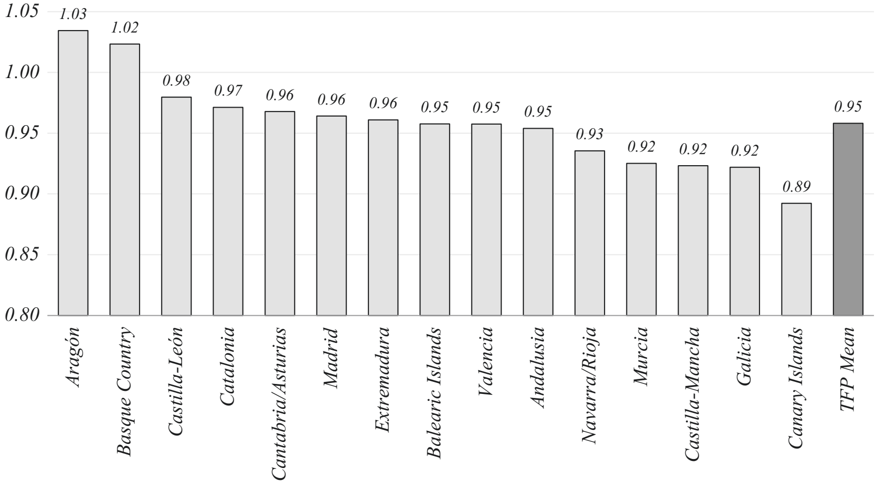


Fig. 2 Regional productivity ranking for Spanish hotels (2001–2008)

second group includes those regions with productivity decline. In this group, Galicia and the Canary Islands are lowest ranked in terms of regional productivity change.

It is worth noting that the growth in *PLAC*, growth in *OCCUP,* and growth in *GDP* in the region of Aragon are above the mean (see Table 3); in particular, growth in *PLAC* is the highest of all the regions in the period 2006–2008. Aragon is also the top-ranked region in terms of the growth in *OCCUP*. The Basque country is also above the national average in terms of tourist flows. Although the growth in GDP in Basque country is slightly below the mean for the periods 2001–2006 and 2001–2008, it is highest ranked for the period 2007–2008.

We use the Kruskal-Wallis test to analyze whether the TFP exhibits statistically significant differences by region. We find statistically significant differences by region using the whole period 2001–2008 ($\chi^2 = 43.708$, *p-value* $= 0.0001$), the initial period 2001–2006 ($\chi^2 = 46.776$, *p-value* $= 0.0001$) and the subsequent period 2007–2008 ($\chi^2 = 25.425$, *p-value* $= 0.0306$). We use the 5% for the confidence level of the Kruskal-Wallis test.

5.2 *Results of the Study 2*

Before presenting the results of the model regression, we explore the correlation matrix between the variables. Table 6 represents the pairwise Pearson (below the diagonal) and Spearman (above the diagonal) correlations. Briefly, TFP is positively related to the change in *OCCUP* (*p-value* $= 0.014$, Pearson; *p-value* $= 0.074$, Spearman) and *GDP* (*p-value* $= 0.092$, Pearson; *p-value* $= 0.093$, Spearman). In contrast, TFP is negatively associated with *PLAC* (*p-value* $= 0.095$, Pearson; *p-value* $= 0.064$

Table 6 Pearson (below diagonal) and Spearman (above diagonal) correlations

	TFP	Change in PLAC	Change in OCCUP	Change in GDP	Change in SIZE	Change in LEV
TFP	—	−0.033 0.064	0.059 0.074	0.062 0.093	−0.322 0.000	−0.104 0.000
Change in PLAC	−0.011 0.095	—	0.190 0.000	0.481 0.000	0.118 0.000	−0.031 0.081
Change in OCCUP	0.044 0.014	0.181 0.000	—	−0.068 0.000	0.049 0.006	0.000 0.980
Change in GDP	0.018 0.092	0.579 0.000	−0.045 0.010	—	0.064 0.000	−0.022 0.213
Change in SIZE	−0.269 0.000	0.121 0.000	0.076 0.000	0.075 0.000	—	0.288 0.000
Change in LEV	−0.077 0.000	−0.031 0.076	−0.019 0.286	−0.021 0.234	0.141 0.000	—

All variables are defined in Table 1

Spearman), *SIZE* (*p-value* < 0.01, Pearson and Spearman) and *LEV* (*p-value* < 0.01, Pearson and Spearman). The negative and significant relationship between TFP and *SIZE* and TFT and *LEV* suggests that changes in Size and changes in leverage affect negatively to TFP. Kim [26] also shows that large hotels and with high leverage hotels are less productive.

Table 7 summarizes the regression of productivity change in regional tourist flow and the firm variables. The first column shows the coefficients of the regressors, the second column shows the standard errors, while the last column shows the two-tail p-values that test the hypothesis that each coefficient is different from 0. In the regression model, we considered three levels of confidence: statistical significance at the 1 and 5% levels, as well as marginal statistical significance at the 10% level.

Most previous research analyzed the effect of macroeconomic variables on hotel revenues and operating results, using different empirical models (see e.g. [12, 21, 35, 41]). The link between TFG in hotels firms and changes in macroeconomic variables has been examined, but to a lesser extent. Table 7 shows that the coefficient on growth in *PLAC* is negative with marginal statistical significance (*p-value* = 0.093), the coefficient on the growth in *OCCUP* is positive and statistically significant (*p-value* = 0.001) and the coefficient on *GDP* is positive with marginal statistical significance (*p-value* = 0.091).

Future research should explore the link between tourism flows and hotel performance. Factors related to visitor arrivals (foreign or domestic, individual or groups, country of origin etc.), rooms (day-stay, guest nights, room nights, etc.), seasonal

Table 7 Results of the panel regressions models

TFP	Coef exp	Coef.	Std. error	$P > \|t\|$
Change in PLAC	–	−0.042	0.072	0.093
Change in OCCUP	+	1.215	0.367	0.001
Change in GDP	+	0.010	0.007	0.091
Change in SIZE	–	−0.044	0.002	0.000
Change in LEV	–	−0.006	0.003	0.080
Cons		−0.104	0.104	0.308
Year fixed effects		YES		
Obs		3,186		
F (7. 3185)		68.72		
Prob > F		0.0000		
R-squared		11%		

All variables are defined in Table 1

effects, and hotel capacity effects could help to better understand the main factors, and moreover the factors that matter most to hotel efficiency.

Size of hotel firms has been extensively analyzed in performance and efficiency research with mixed results. Some studies show that larger hotels are more efficient while other studies show that small hotels also achieve good levels of efficiency. However, studies that consider the TFP of hotel firms are scarce. Table 7 shows that the coefficient on change in *SIZE* is negative and statistically significant (*p-value* = 0.000).

Finally, the coefficient on *LEV* is negative with marginal statistical significance (*p-value* = 0.080). This means that changes in leverage negatively affect TFP. Although this paper includes two factors related to firm structure, future studies should also include factors such as the organizational structure of the firm's production units – both vertical and horizontal [36].

In short, this paper offers an empirical approach to examine the relationship between TFP and firm and regional industry factors. As some variables of the model are significant at the 10% confidence level, preliminary findings from this paper should be sustantiated in future papers.

In sum, this paper offers an empirical approach of the relationship between TFP and firm factors and regional industry factors. As some variables of the model are significant considering the confidence level of 10%, the preliminary evidence obtained in this paper should be reinforced in future papers.

6 Discussion and Conclusions

This study provides evidence of the productivity change in Spanish hotel firms from both industry and regional perspectives. The current paper overcomes some limitations in previous studies of productivity in the hospitality industry. First, most DEA efficiency studies in the hospitality industry focus on the performance of hotels using a cross-sectional sample. Second, most empirical studies use a relatively low number of observations. Third, no previous papers have analyzed regional productivity change with microeconomic panel data in Spanish hotel firms. Finally, prior research has not focused on the use of managerial variables and tourism-specific variables to explain hotel firm productivity change.

In this study, regional productivity change is measured using the Malmquist nonparametric technique, which is calculated using the DEA linear programming approach. The evolution of the total factor productivity change (*TFPCH*) shows a decline not motivated by a change in efficiency (*EFFCH*), which in the period 2001–2008 registered positive growth (6.10%). In this case, the main empirical finding is that the productivity index *(TFP)* registered an average decrease of 4.2%, which is almost completely attributable to the unfavorable trend in technical change (*TECH*). Moreover, the Kruskall Wallis test shows significant differences depending on the region where the firm is located. The evidence suggests that Aragon and the Basque Country are the regions with the highest levels of productivity growth. In contrast, the regions with the lowest levels of efficiency are the Canary Islands and Galicia.

In a second stage, we include two groups of key factors as possible determinants of hotel performance and productivity: regional industry factors and firm factors. The findings indicate that the productivity index is associated with both regional tourist flows and firm factors. In particular, hotel productivity is positively associated with growth in hotel occupancy rates and GDP. In contrast, the hotel productivity index is negatively associated with change in the estimated hotel capacity of a region.

Firm variables also influence on the TFP index. Size tends to produce a negative effect on hotel productivity growth. The fact that the hospitality industry is highly seasonal increases the necessity of maintaining high investments in infrastructure. Hotel infrastructures remain unused during the remainder of the year and this lack of use can be regarded as a significant source of inefficiency and lack of competitiveness. Future papers should include segmentation by hotel size: micro-hotels, small hotels, medium hotels and large hotels, to provide more detailed evidence about the relationship between hotel size (change in size) and performance.

Finally, the tourism economy and the hotel industry in particular are affected by the location of the firm and other regional factors. Because tourist flows are important for the performance of hotel firms, this paper may help policymakers to design strategies for the tourism industry that ensure sustainable development. In particular, it may be of interest to consolidate positioning in regions with lower productivity results by designing business strategies that improve the competitiveness of hotel firms and increase the value added by customer orientation, quality management, product diversification and markets.

References

1. Barros, C.P.: A stochastic cost frontier in the Portuguese hotel industry. Tour. Econ. **10**, 177–192 (2004)
2. Barros, C.P., Alves, F.P.: Productivity in the tourism industry. Int. Adv. Econ. Res. **10**, 215–225 (2004)
3. Barros, C.P.: Measuring efficiency in the hotel sector. Ann. Tour. Res. **32**, 456–477 (2005a)
4. Barros, C.P.: Evaluating the efficiency of a small hotel chain with Malmquist productivity index. Int. J. Tour. Res. **7**, 173–184 (2005b)
5. Barros, C.P., Dieke, P.U.: Measuring the economic efficiency of airports: a Simar-Wilson methodology analysis. Transp. Res. Part E: Logist. Transp. Rev. **44**(6), 1039–1051 (2008)
6. Barros, C.P., Botti, L., Peypoch, N., Solonandrasana, B.: Managerial efficiency and hospitality industry: the Portuguese case. Appl. Econ. **43**, 2895–2905 (2011)
7. Botti, L., Briec, W., Cliquet, G.: Plural forms versus franchise and company-owned systems: a DEA approach of hotel chain performance. Omega **37**, 566–578 (2009)
8. Caves, D.W., Christensen, L.R., Diewert, W.E.: The economic theory of index number and the measurement of input, output and productivity. Econometrica **50**, 1393–1414 (1982)
9. Chang, H., Choy, H.L., Cooper, W.W., Ruefli, T.W.: Using malmquist indexes to measure changes in the productivity and efficiency of US accounting firms before and after the Sarbanes-Oxley Act. Omega **37**, 951–960 (2009)
10. Chen, C.F.: Applying the stochastic frontier approach to measure hotel managerial efficiency in Taiwan. Tour. Manag. **28**, 696–702 (2007)
11. Chen, C.F., Soo, K.T.: Cost structure and productivity growth of the Taiwanese international tourist hotels. Tour. Manag. **28**, 1400–1407 (2007)
12. Chen, M.H.: The economy, tourism growth and corporate performance in the Taiwanese hotel industry. Tour. Manag. **31**, 665–675 (2010)
13. Elfring, T.: The main features and underlying causes of the shift to services. Serv. Ind. J. **9**, 337–356 (1989)
14. Färe, R., Grosskopf, S., Norris, M., Zhang, Z.: Productivity growth, technical progress, and efficiency change in industrialized countries. Am. Econ. Rev. **84**, 66–83 (1994)
15. Färe, R., Grosskopf, S., Roos, P.: Malmquist productivity indexes: a survey of theory and practice. In: Färe, R., Grosskopf, S., Russell, R.R. (eds.) Index Numbers: Essays in Honour of Sten Malmquist. Kluwer Academic Publishers, Boston (1998)
16. Farrell, M.J.: The measurement of productive efficiency, J. R. Stat. Soc. Ser. A **(CXX)**, 253–90 (1957)
17. Fiordelisi, F., Molyneux, P.: Total factor productivity and shareholder returns in banking. Omega **38**, 241–253 (2010)
18. Goncalves, O.: Efficiency and productivity of French ski resorts. Tour. Manag. **36**, 650–657 (2013)
19. Haugland, S.A., Myrtveit, I., Nygaard, A.: Market orientation and performance in the service industry: a data envelopment analysis. J. Bus. Res. **60**, 1191–1197 (2007)
20. Hsieh, L.F., Lin, L.H.: A performance evaluation model for international tourist hotels in Taiwan: an application of the relational network DEA. Int. J. Hosp. Manag. **29**, 14–24 (2010)
21. Huang, Y., Mesak, H., Hsu, M.K., Qu, H.: Dynamic efficiency assessment of the Chinese hotel industry. J. Bus. Res. **65**, 59–67 (2012)
22. Huang, C.W., Ho, F.N., Chiu, Y.H.: Measurement of tourist hotels' productive efficiency, occupancy, and catering service effectiveness using a modified two-stage DEA model in Taiwan. Omega **48**, 49–59 (2014)
23. Hwang, S., Chang, T.: Using data envelopment analysis to measure hotel managerial efficiency change in Taiwan. Tour. Manag. **24**, 357–369 (2003)
24. Johns, N., Wheeler, K.: Productivity and performance measurement and monitoring. Teare. Strateg. Hosp. Manag. 45–71 (1991)
25. Kao, C., Liu, S.T.: Multi-period efficiency measurement in data envelopment analysis: the case of Taiwanese commercial banks. Omega **47**, 90–98 (2014)

26. Kim, E.: The impact of family ownership and capital structures on productivity performance of Korean manufacturing firms: corporate governance and the "chaebol problem". J. Jpn. Int. Econ. **20**, 209–233 (2006)
27. Kim, S.: Factor determinants of total factor productivity growth in the Malaysian hotel industry: a stochastic frontier approach. Cornell Hosp. Q. **52**, 35–47 (2011)
28. Luo, H., Yang, Y., Law, R.: How to achieve a high efficiency level of the hotel industry? Int. J.Contemp. Hosp. Manag. **26**(8), 1140–1161 (2014)
29. Oh, D., Heshmati, A., Lööf, H.: Technical change and total factor productivity growth for Swedish manufacturing and service industries. Appl. Econ. **44**(18), 2373–2391 (2012)
30. Parte, L., Alberca, P.: New insights into dynamic efficiency: the effects of firm factors. Int. J. Contemp. Hosp. Manag. **27**(1), 107–129 (2015a)
31. Parte, L., Alberca, P.: Determinants of technical efficiency in the Spanish hotel industry: regional and corporate performance factors. Curr. Issues Tour. **18**(4), 391–411 (2015b)
32. Pulina, M., Detotto, C., Paba, A.: An investigation into the relationship between size and efficiency of the Italian hospitality sector: a window DEA approach. Eur. J. Oper. Res. **204**(3), 613–620 (2010)
33. Shang, J.K., Wang, F.C., Hung, W.T.: A stochastic DEA study of hotel efficiency. Appl. Econ. **42**(19), 2505–2518 (2010)
34. Shephard, R.W. (ed.): Theory of Cost and Production Functions, Princenton University Press, Princenton (1970)
35. Sun, S., Lu, W.M.: Evaluating the performance of the Taiwanese hotel industry using a weight slacks-based measure. Asia-Pacific J. Oper. Res. **22**(04), 487–512 (2005)
36. Syverson, C.: What determines productivity? J. Econ. Lit. **49**(2), 326–365 (2011)
37. Witt, C.A., Witt, S.F.: Why productivity in the hotel sector is low. Int. J. Contemp. Hosp. Manag. **1**, 28–33 (1989)
38. Wooldridge, J.M.: Introductory econometrics: a modern approach. Thomson South Western, 3rd edn. (2005). ISBN 0-324-28978-2
39. World Tourism Organization: UNWTO Tourism Highlights. www.unwto.org (2014). Accessed 4 July 2014
40. Xu, J., Wei, J., Zhao, D.: Influence of social media on operational efficiency of national scenic spots in china based on three-stage DEA model. Int. J. Inf. Manag. **36**(3), 374–388 (2016)
41. Yang, Z., Xia, L., Zhong, L., Hu, R.: China's regional hotel industry: efficiencies and promotion. Tour. Trib. **30**(5), 31–44 (2015)
42. Yin, P., Tsai, H., Wu, J.: A hotel life cycle model based on bootstrap DEA efficiency: the case of international tourist hotels in Taipei. Int. J. Contemp. Hosp. Manag. **27**(5), 918–937 (2015)
43. Yu, M.M., Chen, L.H.: Productivity growth of taiwanese international tourist hotels in a metafrontier framework. Cornell Hosp. Q. (2015)

Sugarcane Industry in Brazil: Different Impacts for Distinct Municipalities Development Patterns

Márcia Azanha Ferraz Dias de Moraes, Carlos Eduardo Caldarelli and Leandro Gilio

Abstract The increased demand for ethanol in Brazil and the international interest in alternative energy sources and less environmentally harmful fuels stimulated a significant growth in Brazilian sugarcane, sugar and ethanol production, with the expansion of sugarcane agricultural area and new processing units of ethanol and sugar. This Chapter assesses the socioeconomic impacts of this recent sugarcane industry expansion over five years, from 2005 through 2009. For this purpose, a panel data analysis was developed considering socioeconomic impacts on different levels of municipalities' development. The results suggested that sugarcane, sugar and ethanol production can improve socioeconomic indicators, mainly in municipalities that have low and medium level of development, besides the environmental ethanol benefits widely discussed in the literature. These findings indicate that public policies for the sector should consider socioeconomic aspects, both in Brazil as in other developing nations.

Keywords Brazil sugarcane industry · Sugarcane ethanol · Socioeconomic impacts · Quantile regression

M.A.F.D. de Moraes (✉) · L. Gilio
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo,
Avenida Pádua Dias, 11 - Piracicaba/SP, São Paulo 13400-970, Brazil
e-mail: mafdmora@usp.br

L. Gilio
e-mail: lgilio@usp.br

C.E. Caldarelli
Universidade Estadual de Londrina, Rodovia Celso Garcia Cid - PR 445 KM 380 Campus
Universitário-Londrina/PR, P.O. Box 6001, Londrina, Brazil
e-mail: carlos.caldarelli@gmail.com

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_4

1 Introduction

In the post 2000 period, several factors led to an increase in the production of biofuels in countries with productive potential, as Brazil, United States and some European Union countries. Looking at the Brazilian context, where ethanol[1] is made from sugarcane, the production expansion was spurred mainly by the introduction in 2003 of flexible-fuel vehicles (capable of running on any arbitrary combination of gasoline and ethanol), which increased the demand for hydrous ethanol.[2] In addition, the increased international interest in alternative energy sources and less environmentally harmful fuels created expectations of biofuel demand growth around the world, and Brazil could offer a large share of ethanol supply. This context led to a significant increase in the Brazilian production of sugarcane and ethanol, held both by the expansion of the plants already existing in the country, as the installation of new plants ("greenfield" projects), which included investments by domestic capital and foreign companies.[3]

In effect, the sugarcane industry[4] production increased considerably [24, 36]. Between the harvests of 2001–02 and 2012–13, Brazilian sugarcane production rose from 293 to 588 Mt, sugar production from 19 to 38 Mt, and ethanol production increased from 11 million m^3 to 23 million m^3 [37]. According FAO, actually Brazil is the largest sugarcane producer in the world, followed by India, Thailand and Australia; the largest sugar producer and exporter, besides it is the world's the second-largest ethanol producer [11].

This accelerated growth of sugarcane industry verified in Brazil, concomitantly with the increase in corn ethanol production in the United States and the growth

[1] Brazil produces two types of fuel ethanol: anhydrous ethanol, which is mixed with gasoline; and hydrous ethanol, which can be used in flex-fuel automobiles and in automobiles that run exclusively on fuel ethanol (hereafter referred to as "ethanol-powered automobiles").

[2] According to statistics from the Brazilian National Association of Automobile Manufacturers (ANFAVEA), by June of 2005 flex-fuel vehicles already accounted for more than half of all light commercial Otto-cycle vehicles licensed in Brazil. That proportion is in 2014 an impressive 90%, flex-fuel vehicles accounting for over 50% of the national vehicle fleet [24].

[3] Between 2007 and 2009, there were at least seven major transactions involving national processing facilities and international groups, such as: French group Louis Dreyfus Commodities and the Brazilian company Santelisa Vale; Spanish group Abengoa Bioenergy (a subsidiary of Abengoa S.A.) purchased a number of sugarcane processing facilities; the Bermudan company Bunge Limited acquired the Brazilian sugarcane-processing conglomerate Grupo Moema; Shree Renuka Sugars, India's largest sugar refiner, purchased the Brazilian sugar and ethanol producing company Vale do Ivaí, then acquiring the majority share of another such company (Equipav); the largest Brazilian producer of sugarcane, sugar, and ethanol, the Cosan group, became even larger after purchasing the Grupo Nova América, which incorporated an additional milling capacity of approximately 11 million tons; Cosan group later announced a joint venture with Shell International Petroleum Company in 2009. In the 2011–12 harvest, the sugarcane processing capacity of the Cosan group was over 65 million tons. In 2010, there were at least 10 transactions involving the purchase of sugar and ethanol producing facilities in Brazil. For more details see [24].

[4] In Brazil the great majority of productive unit produce both sugar and ethanol from sugarcane, and the available statistics usually are presently jointly for these two sectors. Sugarcane industry in this study refers to three sectors: sugarcane production, sugar plants and ethanol plants.

expectations for the production of biofuels in several EU countries, have raised questions about economic, social and environmental aspects arising from this process. Most of the debate is focused in agricultural land competition between biofuels and food production, known in the literature as food versus fuel debate, besides environmental and social issues that have gained importance in the scientific literature and also for policy makers.

Besides the growth verified in Brazil, several countries began in mid-2000 to promote biofuels through public policies aimed at their adoption, given the potential for mitigation of greenhouse gases of some biofuels compared to fossil fuels. By early 2012, public policies promoting the use of biofuels (production subsidies, transport fuel-tax exemptions, share in total transport fuel obligations), as well as blending mandates, were in place at the national level in at least 46 countries and at the regional level in 26 states and provinces [28]. In addition, fuel-tax exemptions and production subsidies have now been put in place in at least 19 countries [28].

In this context, several studies have highlighted evidences of the possibility of economic growth arising from the sector, reflected on jobs and income creation, which can generate positive net benefits especially for the low-income Brazilian's population [15, 23, 29]. However, there is no consensus on the scientific literature about the impacts led by the production and consumption of agricultural fuels, and several authors have argued about the potential of the negative consequences [12, 13, 21, 33].

Several studies point out the need to offer a comprehensive sustainability assessment regarding biofuels, however it is observed in the literature a relatively limited appraisal on the social and wellness aspects related the growth of biofuel production. As stressed by Talamini et al., environmental, agronomic and technological dimensions were the three primarily discussed areas about sugarcane industry [35]. Chagas et al. also highlighted that socioeconomics impacts are less discussed in the literature and it is verified that the results often presents divergent results [4].

Thus, this study aims to assess the socioeconomic impacts of the expansion of Brazilian sugarcane, sugar and ethanol production in municipalities of the biggest sugarcane producer state (São Paulo), for the period 2005–2009. Socioeconomic impacts were evaluated through use of the Federation of the State of Rio de Janeiro Industries' (FIRJAN) Municipal Development Index (IFDM) as a proxy for the Human Development Index (HDI).

Two empirical approaches are used in this research. First, to measure the mean impacts, a panel data analysis is implemented. Secondly, a quantile regression approach is used in order to measure the socioeconomic impacts considering different municipalities' levels of development [17]. The main innovation of this study is the assessment of the impacts considering the different patterns of development of the municipalities.

2 Literature Review

It is noticed when analyzing the literature about the socioeconomic impacts related to the expansion of sugarcane crops and biofuel production the prevalence of studies related to the impacts on the labor market and rural workers. According to Carvalho and Marin, this recurrence may be justified by the fact that the promotion of access to work constitutes a major mechanism of social inclusion, generally associated with agro energy policies [4].

The jobs created by the sector's presence are those ones directly related to the production of sugarcane, sugar and ethanol sectors (sugarcane industry) as well those ones generated due the interactions of these sectors with other sectors of the economy, whether as a purchaser of the inputs needed for production, as a supplier of products for indirect use, or as a supplier of products for direct use (by the final consumers of sugar and ethanol). However, in the scientific literature about the impacts of sugarcane, sugar and ethanol sectors it is noted a greater attention to jobs creating by the direct way.

Moraes conducts a descriptive analysis of the Brazilian National Household Sample Survey (PNAD) and Annual Social Information Report (RAIS) highlighting the growth of number of employees and formalization of work within the sector [23]. This study indicates that between 2000 and 2005, considering the three sectors of sugarcane industry (sugarcane, sugar and ethanol), there was a significant increase of 52.9% in the number of employees, which increased from 642,848 in 2000 to 982,604 in 2005 [23]. Coelho et al. highlight the low cost of creating a job in the sugarcane agribusiness with respect to chemical and petrochemical industry [6]. According to these authors, a new job in the chemical and petrochemical industry can cost up twenty times more and the employment rate per unit of energy produced is up to 152 times higher in the ethanol agro industry compared with the oil industry (or fossil fuel) [6].

Regards economic development, several authors point out that the expansion of the sugarcane, sugar and ethanol production can contribute for the economic development in rural areas, which usually presents worse socioeconomic indicators compared to urban or industrial areas.

In the literature about the theme, the presence of the sugarcane mills is reported as a key driver of endogenous growth in the municipalities [32]. In addition, to the direct relationship arising from job generation, some authors reports the effects on local business or services, urbanization, income, population expansion and growth of municipal tax collection [5, 22, 26, 29, 32].

Regarding the aspect of aggregate income, studies about the sector's expansion in general are convergent in affirming the positive impacts (some of small magnitude) of the sector's presence. Oliveira et al. assessed whether the expansion of sugarcane in the Midwest of the State of Minas Gerais, which intensified after 2006, has contributed to higher growth in GDP per capita of the municipalities of this Brazilian State [26]. This study reports that in the municipalities evaluated from 1999 to 2008 period, the growth was 39.94% where it occurs with the presence of the sugarcane industry, while the average of GDP growth was 22.49% for the Midwest of Minas Gerais, and 29.1% for the state as a whole [26]. However, these data must be evaluated carefully, as that study does not include other regional factors that may bias the analysis.

Among the studies of quasi-experimental approach, Deuss used a propensity score matching (PSM) method to evaluate the effect of the expansion of the sugarcane industry (treatment effect) on the brazilian economic development at the municipality level in Brazil as a whole and in the main sugarcane producing regions, the North-Northeast (NE) and the Center-South (CS) [7]. As a result, the author found a positive GDP per capita effect, especially in the South Central region (except São Paulo) and Northeast. That study did not find a significant effect on economic growth in the State of São Paulo [7].

Satolo and Bacchi analyzed the impact of sugarcane and ethanol expansion in the state of São Paulo, assessing their impact on GDP per capita of different municipalities [29]. Through a spatial dynamic panel data model, the study shows that there is a positive spatial time dependence on the level of per capita GDP and on its distribution [29]. The authors found that the effect of the expansion of sugarcane industry is positive on GDP per capita if this expansion occurs in an area of up to 23% of the municipalities' agricultural areas, replacing crops or pasture areas [29]. This study, by aggregating spatial analysis, also evaluates a positive impact on the sugarcane industry's presence on the nearby municipalities, although it was a small effect. This spillover effect can be explained by migratory attraction and increased own local income, which can increase demand for goods and services consumed locally, multiplying the positive effect on income.

Bacchi and Caldarelli undertook a panel data analysis in order to identify the positive externalities related to the expansion of the sugarcane industry by evaluating the Municipal Development FIRJAN Index (IFDM). These authors present evidence that the expansion of the sugarcane industry in the state of São Paulo generated positive effects on employment and income, but there was no significant positive impacts on health and education indexes. Nevertheless, the authors didn't analyze the differences between less developed municipalities and more developed ones [2].

In addition to these aspects, Shikida and Souza argued that the presence of plants and sugarcane plantations contributes to smooth the evasion of rural people of the municipalities, which may occur with the decrease of family farming areas [32].

In contrast with benefits evaluated in these studies, Sawyer points out a possible effect of concentration of income, driven by the expansion of a culture over large areas [30]. The author argues that both in the case of manual harvesting sugarcane, with working conditions often precarious, as the mechanized, extinguishing jobs, there may be this negative impact locally [30].

With respect to reduction of regional inequalities, Schaffel and Rovere estimate that the expansion of the sugarcane industry has had little influence [31]. This fact is justified, according to the authors, because the production of ethanol and sugar are concentrated mostly in São Paulo, state with high level of development. However the authors point out that the expansion into new areas is still in the beginning, and there are no several impact studies [31].

Chagas, Toneto-Jr and Azzoni sought to identify the effects that the production of sugarcane has on the social indicators of the producing regions using municipal Human Development Index (HDI) as a summary indicator of local social conditions [8]. Through the spatial propensity score matching method, these authors concluded that the sugarcane industry's presence in the evaluated municipalities is not relevant to determining their social conditions, for better or for worse [5].

In different case studies it is also reported the precariousness of jobs created, the risks to health of rural workers and poor housing conditions of immigrant workers, which suggests the occurrence of negative impacts at the local level [12, 21].

The findings of the literature review indicates a predominance of the case studies analysis, which are important for the analysis and understanding of particular realities in detail, but are limited in characterizing the impacts of the sugarcane industry expansion in a more comprehensive and broader way, due particular institutional, economic and social characteristics of different municipalities, regions or countries.

It is also interesting to observe the divergent results between qualitative and quantitative analysis found in the literature review, especially under different geographical levels. The quantitative approach studies; using econometric methods, general equilibrium or critical analysis of data, tend to have a more positive outlook when compared to the case studies.

Despite the importance of the cited studies and apparent contradictions in the results, it can be seen in the literature a relative shortage of research on evaluation and assessment of social impacts of sugarcane ethanol expansion, in contrast to the further investigation of the environmental and agronomic aspects.

Therefore, the empirical analysis developed and described in the following sections of this chapter seeks to contribute to the evaluation of the relationship between the expansion of the sugarcane industry and socioeconomic development patterns, analyzing the effects of the recent sugarcane expansion and the presence of plants on producing municipalities of the State of São Paulo, the main Brazilian producer state.

More particularly, this study assesses the socioeconomic effect through the Municipal Development FIRJAN Index (IFDM), a summary index that annually reports the socioeconomic development of municipalities [10]. The period of analysis is 2005–2009, and the emphasis of this empirical study is on analyzing socioeconomics impacts considering municipalities different' levels of development, given the lack of studies with this approach.

3 Methodological Procedures

We address the issue of the effects of sugarcane expansion on socioeconomic development of São Paulo municipalities' in two different methodological approaches. First, we conduct a panel data analysis according to steps proposed by Greene [14] to measure the mean impacts, taking into account differences in behavior across individuals. Second, in order to measure the socioeconomic impacts considering different municipalities' levels of development, a quantile regression approach is adopted [17].

3.1 Panel Data Analysis and Quantile Regression

Panel data, also known as longitudinal or cross-sectional time series data, is a data set in which the behavior of individuals/units is observed across time. Data sets that combine time series and cross-section are common; these kinds of datasets provide a rich source of information [14].

This study uses a panel data analysis because it allows measuring the socioeconomic impacts of the sugarcane industry in municipalities across time. According to Maddala, the methodology takes into account heterogeneity across units, the analysis allows to control for variables that change over time but not across individuals/units (national policies, federal regulations, international agreements) [19]. So, there is a great flexibility in modeling differences across individuals.

The basic framework for i units and t periods is a regression model [14, 19]:

$$y_{it} = x_{it}'\beta + z_i'\alpha + \varepsilon_{it}, \tag{1}$$

where there are k regressors in x_{it} and the main objective of the analysis will be consistent and efficient estimation of the partial effects (β),

$$\beta = \partial E[y_{it}|x_{it}]/\partial x_{it}. \tag{2}$$

The heterogeneity is $z_i'\alpha$ where z_i contains a set of individual or group specific variable which may be observed or sometimes unobserved – are the set of missing variables. There are different kinds of panel data structures; which depend on the missing variables z_i, that is:

- **Pooled regression** – if z_i contains only a constant term, there is a common α;
- **Fixed Effects** – if the z_i is unobserved and correlated with x_{it};
- **Random Effects** – if the z_i is unobserved and uncorrelated with x_{it}.

The Fixed Effects model are used whenever you are only interested in analyzing the impact of variables that vary over time, and the Random Effects model assume that the entity's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables. In random-effects you need to specify those individual characteristics that may or may not influence the predictor variables [34]. Some tests are performed to decide which model fits better and shall be estimate, as Hausman, Breusch and Pagan and Chow test.[5]

[5] For more details see [26, 32, 34].

Quantile regression, as Koenker and Basset defines it, is a method for estimating functional relations between variables for all portions of the probability distribution – different quantiles (τ) [17].

As described by Koenker, quantile regression models present many new possibilities for statistical analysis and interpretation of economic data, mainly because this analysis allows comparing how some percentiles may be more affected by certain characteristics than other quantiles [18]. This is reflected on the size change of the regression coefficient.

The conditional quantile is denoted by:

$$Qy_{it}(\tau|x'_{it}) = x'_{it}\beta(\tau) + z'_i\alpha(\tau). \quad (3)$$

For this study, we consider that z_i contains only a constant term.

The advantage of using quantile regression to modeling the socioeconomic impacts related to the existence of sugarcane industry in the municipality is the possibility to compare these impacts according to the different levels of development of the municipalities for the State of São Paulo.

3.2 Data and Empirical Strategy

In order to measure the socioeconomic impacts of the sugarcane industry in the municipalities of the state of São Paulo, we estimate the proposed model using panel data analysis methodology and quantile regression approach:

$$INDEX_{it} = [\text{DU, area, GDP per capita}]'_{it}\beta + z'_i\alpha + \varepsilon_{it}, \quad (4)$$

where:

- **INDEX** is the FIRJAN Municipal Development Index (IFDM) used to measure the level of development in each municipality of the state of São Paulo – the index varies from 0 (least developed) to 1 (most developed);
- **DU** is a dummy variable used to identify the existence of sugar and/or ethanol mills/distillery in each municipality – 1 if it has a mills/distillery and 0 if it hasn't;
- **area** represents the sugarcane harvest area in each municipality (the variable is the percentage of the total area of agriculture, cattle and pasture in each municipality), and;
- **GDP per capita** is used because it is well known that other factors could influence the Firjan development index, and the GPD is an important variable to capture this effect – it is a control variable – in 2008 US$.

Furthermore, we use binaries variables for years; the objective is control the time effect as suggested by Greene [14].

The information used to build the database were collected from IBGE (Brazilian Institute of Geography and Statistics), FIRJAN (Rio de Janeiro Federation

Table 1 Description of the IFDM indicators and components [16]

IFDM		
Employment and income[a]	Education[b]	Health[c]
• Formal jobs	• Primary school enrollment	• Number of prenatal consultation
• Jobs for local workers	• Primary school leaver	• Death due to not defined cause
• Formal income generating	• Age-series distortion on primary school	• Child mortality
• Median wages	• Undergraduate teachers in primary school	• Hospitalizations
• Income inequality	• Average hours in class	
	• IDEB index[d]	

Source Performed by authors

[a] Data from Brazilian Ministry of Labor and employment (MTE)

[b] Data from Brazilian Ministry of Education (MEC)

[c] Data from Brazilian Ministry of Health (MS)

[d] The IDEB is the Basic Education Development Index carried out by Brazilian Ministry of Education to evaluate the school quality and the students' performance

of Industries), IPEA (Applied Economic Research Institute) and UNICA (Brazilian Sugarcane Industry Association) [3, 10, 16, 36].

To verify the existence of sugar and/or ethanol mills/distillery for each municipality the information from UNICA [36] are used. The sugarcane area (hectare), crops, pasture and cattle were obtained from IBGE data base (SIDRA/IBGE) [16]. To build the GDP per capita series we use data from IBGE and IPEA [3, 16]. Finally, we use data from FIRJAN for the development index [10]. The FIRJAN Municipal Development Index (IFDM) closely follows the annual social and economic development of municipalities, reporting on employment and income, education and health issues. The IFDM follows the IDH (ONU) methodology. Table 1 shows the IFDM composition.

The study was realized with annual data for the period 2005–2009 using the commercial statistical package STATA® 10.0. The study was performed using data for the São Paulo state; which is the main sugarcane, sugar and ethanol producer in Brazil. The models were estimated using aggregate IFDM index and sub-index IFDM Employment and Income.

4 Results and Discussions

We can observe important changes in land use as a consequence of the sugarcane industry's expansion in Brazil. First, the expansion is concentrated in the center-south region, especially in the states of São Paulo, Mato Grosso do Sul, Mato Grosso

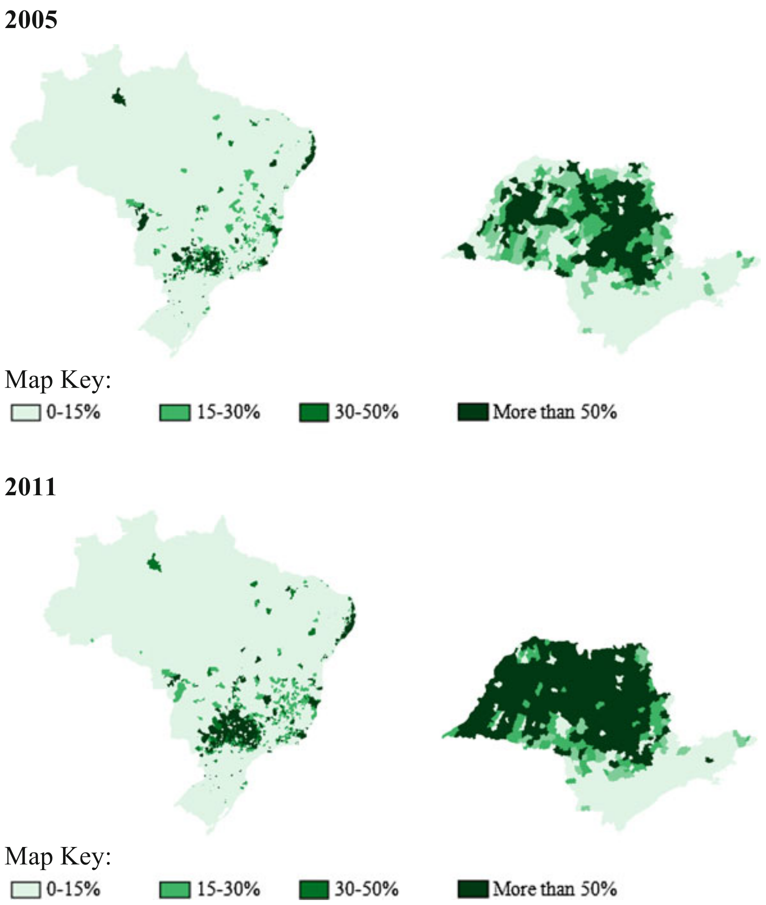


Fig. 1 Sugarcane area in Brazil and in the State of São Paulo for 2005 and 2012 – Percentage of agricultural area allocated to sugarcane crops [16]. *Source* performed by authors

and Minas Gerais. In the case of the state of São Paulo, 1990–91 harvest accounts for 59.26% of the Brazilian sugarcane production; in 2001–02 the percentage was 60.26% and in 2012–13 this participation was 56.06% [36]. Figure 1 shows the evolution of sugarcane cultivation in Brazil and in the State of São Paulo.

Figure 2 presents the Aggregated IFDM and the Employment and Income IFDM evolution for all the São Paulo municipalities, from 2005 to 2009. The overall analysis show that Employment and Income indicador has a more dispersed distribution than Aggregated IFDM. The characteristic of these data corroborate the methodological tools used.

It is also important to underline that both indicators have presented constant trend for the analysed period, excepted for 2008, when the Brazilian economy was impacted for the international crisis. According to Paula and Ferrari-Filho the most intense

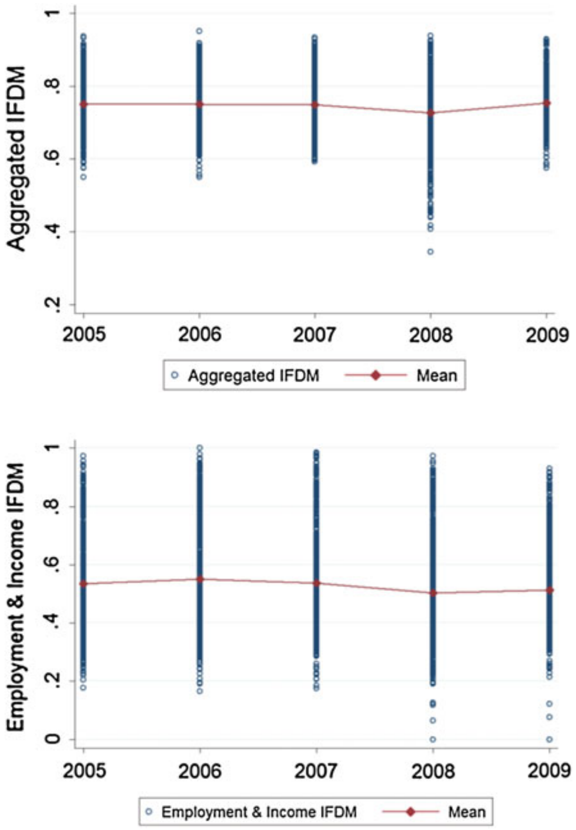


Fig. 2 Evolution of the IFDM index – Aggregated IFDM and Employment and Income IFDM – for all São Paulo municipalities from 2005 to 2009. *Source* Performed by authors using FIRJAN [10]

impacts of the international crisis in the Brazilian economy were observed in the second semester of 2008; the most affected economic variables were job market and GDP, therefore the observed decrease on the mean of the Employment and Income IFDM in 2008 [27].

Table 2 presents the variables' descriptive statistics. We considered 3225 observations for each variable. The observations contain data from all 645 municipalities in the Brazilian state of São Paulo over 5 years, 2005 through 2009. Since this is a panel data analysis, the descriptive statistics presented in Table 2 are divided into the dimensions Within (variance between municipalities) and Between (average variation over a period of time).

Table 3 presents the results of the estimatives of the socioeconomic impacts of sugarcane industry using panel data analysis.

The findings (Table 3) suggest that there is a positive and statistically significant impact between the explanatories variables sugar mills and/or ethanol distilleries (DU) and area of sugarcane in each municipality (area) on the dependent variable development indexes (IFDM and IFDM Employment and Income).

Table 2 Summary statistics for the variables used in the model estimation

Variables	Mean	s.d	Min.	Max.	Obs.
IFDM *overall*	0.75	0.06	0.54	0.95	N = 3225
between		0.06	0.60	0.92	n = 645
within		0.02	0.61	0.86	T = 5
IFDM_ER *overall*	0.53	0.16	0	1	N = 3225
between		0.15	0.28	0.95	n = 645
within		0.07	0.01	0.83	T = 5
area *overall*	0.18	0.14	0	0.79	N = 3225
between		0.14	0	0.72	n = 645
within		0.12	0.01	0.74	T = 5
GDP per capita *overall*	6392	601	1214	8768	N = 3225
between		5548	1789	67934	n = 645
within		2324	–28096	38867	T = 5

Source Performed by authors

Table 3 Socioeconomic impacts of the sugarcane industry in the State of São Paulo – from 2005 to 2009 – using panel data analysis (Fixed Effects The fixed effects model was chosen based on Chow, Breusch-Pagan and Hausman tests results)

	Aggregated IFDM			Employment and Income IFDM		
Variable[a]	Coefficient (%)	*t* test	$P > \|t\|$	Coefficient (%)	*t* test	$P > \|t\|$
DU	1.34	2.53	0.01	6.07	3.21	0.00
area	0.30	3.07	0.02	1.22	3.69	0.00
GDP per capita	3.19	2.15	0.03	12.04	3.84	0.00

Source Performed by authors

[a]*Note* Binaries variables for years and units were used to estimate the model, but the results were omitted; Robust standard errors were used

When we analyze the variable DU, we can observe two important results: (i) municipalities that have mills or distilleries have an aggregated index of development 1.34% higher than municipalities that doesn't have; (ii) municipalities that have mills or distilleries presents the index of employment and income about 6.07% higher.

The impacts of the variable area on the development indexes are also positive, the coefficients are respectively 0.30% for IFDM and 1.22% for IFDM Employment and Income; although the coefficients are small when compared to the variable DU.

It is interesting to highlight that development indexes have been widely influenced by the existence of processing sugarcane plants (ethanol and/or sugar plants) than sugarcane production. In general, it was possible to identify a closer relation between

Table 4 Socioeconomic impacts of the sugarcane industry in the State of São Paulo on Aggregated IFDM – from 2005 to 2009 – using quantile regression

Quantile	Variable[a]	Coefficient (%)	t test[b]	$P > \|t\|$
0.25	*DU*	**1.76**	5.30	0.00
	Area	**0.71**	3.66	0.00
	GDP per capita	7.13	15.14	0.00
0.5	*DU*	**1.40**	7.90	0.00
	Area	**0.64**	2.47	0.00
	GDP per capita	8.72	16.38	0.00
0.75	*DU*	**0.93**	3.66	0.00
	Area	**0.45**	2.19	0.02
	GDP per capita	11.02	17.91	0.00

Source Performed by authors

[a] Binaries variables for years were used to estimate the model, but the results were omitted; Robust standard errors were used

[b] Additional tests were performed to evaluate the regressions; the regressions have a good explanatory power

Table 5 Socioeconomic impacts on Employment and Income IFDM index of the sugarcane industry in the State of São Paulo – from 2005 to 2009 – using quantile regression

Quantile	Variable[a]	Coefficient (%)	t test[b]	$P > \|t\|$
0.25	*DU*	**6.84**	7.32	0.00
	Area	**2.11**	3.21	0.00
	GDP per capita	20.00	12.41	0.00
0.5	*DU*	**6.84**	8.31	0.00
	Area	**3.24**	5.54	0.00
	GDP per capita	23.55	18.57	0.00
0.75	*DU*	**3.69**	4.63	0.00
	Area	**2.30**	3.77	0.00
	GDP per capita	32.06	12.82	0.00

Source Performed by authors

[a] Binaries variables for years were used to estimate the model, but the results were omitted; Robust standard errors were used

[b] Additional tests were performed to evaluate the regressions; the regressions have a good explanatory power

sugarcane production/processing and economic development in the State of São Paulo. The results suggest an association of the sugarcane industry in São Paulo with economic development, especially related to the improvement on the job market and income.

In addition, we present the estimative using quantile regression (Tables 4 and 5). In this point, we are interested in better understanding the impacts of the sugarcane industry on different levels of development municipalities.

The results presented on Tables 4 and 5 corroborate the previous analysis using panel data (Fixed Effects Model); the findings also indicate that sugarcane processing (DU) has a higher impact on economic development indexes than sugarcane area (area) – for all estimated quantiles. On the other hand, the above results (Table 5) suggest an important point, the fact of the municipalities with low and medium levels of development can be more impacted by sugarcane industry than the higher developed.

For the less developed municipalities – quantile 0.25 –, the presence of a sugarcane mill or ethanol distillery increases the Aggregated IFDM by 1.76% and the Employment and Income IFDM by 6.84%. For the other extreme of distribution, highest developed municipalities – quantile 0.75 –, the coefficients are smaller, respectively, 0.93 and 3.69%.

These results are corroborated using interquartile regression – interquartile range; as stated in Tables 6 and 7 (Appendix), the negative coefficients for DU variable means that the presence of the sugar or ethanol mills or distilleries decreases the interquartile range and therefore Aggregated and Employment and Income IFDM dispersion, consequently we have expected a downward trend.

Thus, according to the results, the presence of sugarcane facilities can contribute to the convergence of the poorer municipalities' income towards the more developed municipalities. These results are confirmed by Table 8 (Appendix), which shows that it is possible to reject the null hypothesis (that the coefficients of quantiles regression are the same), confirming the differential impact between municipalities according to GDP per capita levels.

We present (Figs. 3 and 4) the pattern of DU variable obtained using alternative methodologies (quantile regression, Ordinary Least Squared and Fixed Effects) for all quantiles to compare the impacts of the presence of sugarcane mills or distilleries in municipalities with different levels of development.

The relation between processing sugarcane (DU) and economic development can be observed in Fig. 3 (Aggregated IFDM) and Fig. 4 (Employment and Income IFDM), according the data distribution – quantiles describing levels of development. The analysis reinforces that presence of the sugarcane mills or distilleries in the municipalities with low or medium level of development may be associated to the highest socioeconomic impacts.

This way, the results show that the presence of sugarcane industry (sugarcane, sugar and ethanol production) improved the development index for the municipalities of the State of São Paulo in the analyzed period. Those findings point out that public fuel policies, in order to expand sugarcane ethanol production, besides having an important environmental benefit can also improve socioeconomic indicators in municipalities that have low and medium level of development.

Fig. 3 Comparing the coefficient DU from different models by quantile – for Aggregated IFDM. *Source* Performed by authors

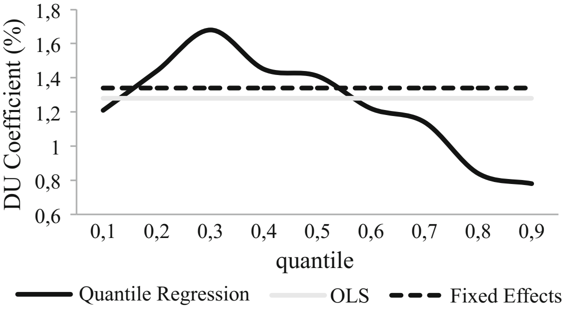
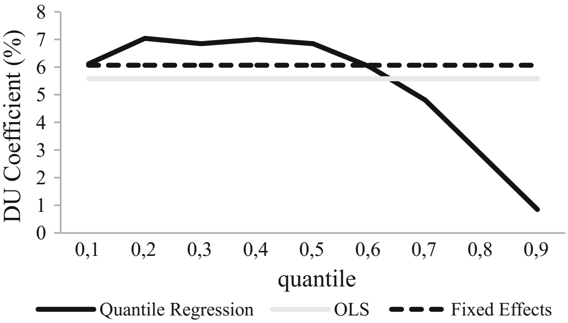


Fig. 4 Comparing the coefficient DU from different models by quantile – for Employment and Income IFDM. *Source* Performed by authors



5 Final Remarks

The replacement of fossil fuels by biofuels in several countries, driven mainly by environmental and energy security concerns, has given rise to discussion about the economic, environmental and social impacts, considering the possibility of sharp growth in biofuels production. On the one hand, academic studies and papers regarding the environmental impacts and on land use flourished, however the social impacts analysis are relatively scarce in the international literature. Moreover, the results on social impacts are quite diverse, which motivated the development of this chapter.

The main innovation of this chapter when comparing with the existing literature is the impacts evaluation considering the different patterns of development of the municipalities. The availability of socioeconomic data series allows the analyses of Brazilian experience under several different approaches, what can be useful for national stakeholders (researchers, police makers, producers), as for the new producer countries that aim to start or expand biofuels production, and aims to assess the socioeconomics impacts.

The approach we use - panel date analysis - has shown that there is a positive and statistically significant effect between the existence of both sugarcane processing plants (sugar mills/ethanol distilleries) and the sugarcane area on the Development Indexes (Aggregated Development Index – IFDM, and Employment and Income Index IFDM).

The results show that municipalities with sugar/ethanol plant and sugarcane area have both IFDM indexes higher than municipalities that don't have. It is interesting to highlight that the impact on the development indexes have been widely higher due to the existence of processing sugarcane plants (ethanol and/or sugar plants) than sugarcane production.

Considering the results using quantile regression, that better allows to address the impacts of the sugarcane industry on different levels of development municipalities, the results also indicate that municipalities that have sugarcane area and sugarcane processing units have better development indexes than those ones that do not have, for all quantiles analyzed. One important finding is that municipalities with low and medium levels of development are more impacted by industry than the higher developed ones. This way, the conclusion is that the sugarcane industry improved the regional development index for the municipalities of the state of São Paulo in the analyzed period.

Those findings point out that besides have an important environmental benefit, as widely discussed in the literature, sugarcane ethanol production can also improve socio-economic indicators, mainly in municipalities that have low and medium level of development.

We hope this research can contribute for national policymakers, as well as for other developing countries that aim to expand biofuels production as well as to improve regional development.

Appendix

See Table 8.

Table 6 Interquartile regression on Aggregated IFDM – from 2005 to 2009

Interquartile range	Variables[a]	Coefficient (%)	*t* test	$P > \|t\|$
0.25 0.75	*DU*	–0.81	–2.83	0.00
	Area	–0.26	–0.71	0.47
	GDP per capita	3.89	7.04	0.00
0.25 0.50	*DU*	–0.35	–1.66	0.09
	Area	–0.06	–0.33	0.74
	GDP per capita	1.58	3.31	0.00
0.50 0.75	*DU*	–0.46	–2.46	0.01
	Area	–0.19	–1.05	0.29
	GDP per capita	2.30	5.34	0.00

Source Performed by authors
[a]*Note* Binaries variables for years were used to estimate the model. but the results were omitted

Table 7 Interquartile regression on Employment and Income IFDM – from 2005 to 2009

Interquartile range	Variables[a]	Coefficient(%)	t test	$P > \|t\|$
0.25 0.75	*DU*	–2.93	–3.07	0.00
	Area	–0.19	0.22	0.82
	GDP per capita	12.06	6.46	0.00
0.25 0.50	*DU*	0.003	0.00	0.99
	Area	1.13	2.22	0.02
	GDP per capita	3.55	2.04	0.04
0.50 0.75	*DU*	–2.93	–4.22	0.00
	Area	–0.94	–1.83	0.06
	GDP per capita	8.50	4.74	0.00

Source Performed by authors

[a]*Note* Binaries variables for years were used to estimate the model. but the results were omitted

Table 8 Joint F-test for equality of different quantiles for DU variable

Hypothesis	Aggregated IFDM	Income and Employment IFDM
test $H_0[q.25 = q.50 = q.75]$	$F(2.3217) = 3.21$	$F(1.3217) = 8.11$
	$Prob > F = 0.04$	$Prob > F = 0.00$
test $H_0[q.25 = q.50]$	$F(1.3217) = 3.26$	$F(1.3217) = 0.00$
	$Prob > F = 0.07$	$Prob > F = 0.99$
test $H_0[q.50 = q.75]$	$F(1.3217) = 4.73$	$F(1.3217) = 15.82$
	$Prob > F = 0.02$	$Prob > F = 0.00$
test $H_0[q.25 = q.75]$	$F(1.3217) = 6.37$	$F(1.3217) = 7.09$
	$Prob > F = 0.01$	$Prob > F = 0.00$

Source Performed by authors

References

1. Agência Nacional De Petróleo, Gás Natural e Biocombustíveis - ANP. Anuário Estatístico Brasileiro do Petróleo, Gás Natural e Biocombustíveis 2013. Brasília: Ministério de Minas e Energia (2013), p. 232
2. Bacchi, M.R.P., Caldarelli, C.E.: Impactos socioeconômicos da expansão do setor sucroenergético no Estado de São Paulo entre 2005 e 2009. Nova Economia, **25**(1), 209–224 (2015)
3. Base de Dados do Instituto de Pesquisas Econômicas Aplicadas – IPEADATA. Dados Macroeconômicos. (Avaliable via DIALOG, 2014). http://www.ipeadata.gov.br/ipeaweb.dll/ipeadata?122363439 (2014). Accessed 20 May 2014
4. Carvalho, S.P., Marin, O.B.: Agricultura familiar e agroindústria canavieira: impasses sociais. Revista de Economia e Sociologia Rural **49**(3), 681–707 (2011)
5. Chagas, A.L.S., Toneto-Jr, R., Azzoni, C.R.: A spatial propensity score matching evaluation of the social impacts of sugarcane growing on municipalities in Brazil. Int. Reg. Sci. Rev. **35**(1), 48–69 (2012)
6. Coelho, S.T., Goldemberg, J., Lucon, O., Guardabassi, P.: Brazilian sugarcane ethanol: lessons learned. Energy Sustain. Dev. **10**, 26–39 (2006)
7. Deuss, A.: The economic growth impacts of sugarcane expansion in Brazil: an inter-regional analysis. J. Agric. Econ. **63**(3), 528–551 (2012)

8. Egeskog, A., Berndes, G., Freitas, F., Gustafsson, S., Sparovek, G.: Integrating bioenergy and food production: a case study of combined ethanol and dairy production in Pontal. Braz. Energy Sustain. Dev. **15**(1), 8–16 (2011)
9. FAO. Sugarcane potentials. 41p. (Avaliable via DIALOG, 2014). http://www.fao.org/ag/AGL/agll/gaez/ds/ds.htm. Accessed 16 Oct 2014
10. FIRJAN. Índice FIRJAM de desenvolvimento municipal – IFDM. Avaliable via DIALOG, 2014). http://www.firjan.org.br/ifdm/. Accessed 20 Mar 2014
11. Food Agriculture Organization – FAO. Data base. (Avaliable via DIALOG, 2014). http://www.faostat.fao.org. Accessed 20 July (2014)
12. Galiano, A.D.M., Vettorassi, A., Navarro, V.L.: Trabalho, saúde e migração nos canaviais da região de Ribeirão Preto (SP), Brasil: o que percebem e sentem os jovens trabalhadores? Revista Brasileira de Saúde Ocupacional **37**(125), 51–64 (2012)
13. Gonçalves, R.J.A.F., Rogrigues, M., Mendonça, R.: Modernização energética e desenvolvimento do setor sucroalcooleiro: reestruturação produtiva do capital e precarização do trabalho nas áreas de cerrado. Revista Percurso **2**(1), 53–72 (2010)
14. Greene, W.: Econometric Analysis, 6th edn, p. 1178. Prentice Hall, New Jersey (2008)
15. Hoffmann, R.: Segurança alimentar e a produção de etanol no Brasil. Revis. Segurança Alimentar e Nutricional **13**, 1–5 (2006)
16. Instituto Brasileiro De Geografia E Estatística (IBGE). Pesquisa Agrícola Municipal – Culturas temporátias e permanentes. Banco de dados agregados: sistema IBGE de recuperação automática – SIDRA. (Avaliable via DIALOG, 2014). http://www.sidra.ibge.gov.br/bda/pesquisas/pam/default.asp?o=18&i=P. Accessed 16 Out 2014
17. Koenker, R., Basset, G.: Regression Quantiles. Econometrica **46**(1), 33–50 (1978)
18. Koenker, R.: Quantile Regression, p. 349. University Press, Cambridge (2005)
19. Maddala, K.L.: Introduction to Econometrics, 4th edn, p. 654. Wiley, New Jersey (2009)
20. Mangoyana, R.B., Smith, T.F., Simpson, R.: A systems approach to evaluating sustainability of biofuel systems. Renew. Sustain. Energy Rev. **25**, 371–380 (2013)
21. Martinelli, L.A., Filoso, S.: Expansion of sugarcane ethanol production in brazil: environmental and social challenges. Ecol. Appl. **18**(4), 885–898 (2008)
22. Martinez, S.H., Eijck, J.V., Cunha, M.P., Guilhoto, J.J.M.: Analysis of socio-economic impacts of sustainable sugarcane-ethanol production by means of inter-regional Input-Output analysis: Demonstrated for Northeast Brazil. Renew. Sustain. Energy Rev. **28**, 290–316 (2013)
23. Moraes, M.A.F.D.D.: O mercado de trabalho da agroindústria canavieira: desafios e oportunidades. Economia Aplicada **11**(4), 605–619 (2007)
24. Moraes, M.A.F.D., Zilberman, D.: Production Of Ethanol From Sugarcane In Brazil. Springer, New York (2014)
25. Nardy, V., Gurgel, A.C.: Impactos da liberalização do comércio de etanol entre Brasil e Estados Unidos sobre o uso da terra e emissão de CO_2. Nova Economia **23**(3), 693–726 (2013)
26. Oliveira, E.G., Ferreira, M.E., Araújo, F.M.: Diagnóstico do uso da terra na região Centro-Oeste de Minas Gerais, Brasil: a renovação da paisagem pela cana-de-açúcar e seus impactos socioambientais. Sociedade & Natureza **24**(3), 545–556 (2012)
27. Paula, L.F., Ferrari-Filho, F.: Desdobramentos da crise financeira internacional. Rev. Econ. Polit. **31**(2), 315–335 (2011)
28. REN.: Renewables 2012 Global Status Report, p. 172, REN21 Secretariat, Paris (2012)
29. Satolo, L.F., Bacchi, M.R.P.: Impacts of the recent expansion of the sugarcane sector on municipal per capita income in são paulo state. ISRN Econ. **2013**, 1–14 (2013)
30. Sawyer, D.: Climate change, biofuels and eco-social impacts in the Brazilian Amazon and Cerrado. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363**(1498), 1747–1752 (2008)
31. Schaffel, S.B., Rovere, E.L.: The quest for eco-social efficiency in biofuels production in Brazil. J. Clean. Prod. **18**(17), 1663–1670 (2010)
32. Shikida, A., Souza, E.C.: Agroindústria canavieira e crescimento econômico local. Revista de Economia e Sociologia Rural **47**(3), 569–600 (2009)
33. Silva, M.A.M.: Produção de alimentos e agrocombustíveis no contexto da nova divisão mundial do trabalho. Revis. Pegada **9**(1), 63–80 (2008)

34. Stock, J., Watson, M.: Introduction to Econometrics, 2nd edn, p. 840. Pearson, São Paulo (2007)
35. Talamini, E., Caldarelli, C.E., Wubben, E.F.M., Dewes, H.: The composition and the impacto of stakeholders' agenda on U.S ethanol production. Energy Policy **50**(1), 647–658 (2012)
36. União Da Indústria De Cana-De-Açúcar – ÚNICA. Database. (Avaliable via DIALOG, 2014). http://www.unicadata.com.br/. Accessed 20 June 2014
37. Wilkinson, J., Herrera, S.: Biofuels in Brazil: debates and impacts. J. Peasant Stud. **37**(4), 749–768 (2010)

The Coordination of Agricultural R&D in the U.S. and Germany: Markets Versus Networks

Barbara Brandl and Katrin Paula

Abstract Making money out of knowledge is a more difficult venture than it might seem due to defining characteristics of knowledge: non-rivalry and non-excludability in consumption. We argue that institutional attempts to overcome this difficulty in knowledge commodification shape the type of technological innovation in an economy. We suggest that two coordination types of R&D can be found: coordination by the market and coordination by networks. Empirically, our analysis is based on a mixed methods approach. We combine qualitative interviews with employees of seed companies in the U.S. and Germany, historical records, and descriptive quantitative analysis of yield developments in several crops. Finally, we compare market concentration in the U.S. and Germany. Our results indicate that coordination of agricultural R&D by the market (as in the U.S. since the 1980s) fosters innovations that are based on explicit knowledge. Furthermore, coordination by the market privileges large companies, tends to lead to a strong market concentration, and limits the development efforts on a few commercially beneficial crops. Coordination of agricultural R&D by networks (as in Germany), on the other hand, fosters innovations that are based on implicit knowledge and privileges medium-sized handcraft-based companies, which maintain innovation activities in a larger spectrum of crops. We conclude that the ban of transgenic seed in Europe cannot only be explained by the

This paper is a revised and shortened version of our paper "Spielarten des Wissenskapitalismus. Die Kommodifizierng von Saatgut in USA und Deutschland", published in Leviathan 2014, 42(4): 539–572.

B. Brandl (✉)
Rachel Carson Center, Ludwig Maximilian University of Munich,
Munich, Germany
e-mail: barbara.brandl@soziologie.uni-muenchen.de

K. Paula
Graduate School of Economic & Social Sciences (GESS),
University of Mannheim, Mannheim, Germany
e-mail: Katrin.Paula@uni-mannheim.de

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_5

consumer protest but might also root in the institutional structure that coordinates agricultural R&D.

Keywords Agricutural innovation · Seed markets · Transgenic seed · Market concentartion

1 Introduction

Research and development (R&D), especially in the field of agriculture, was always of paramount importance for industrial nations. The political regulation of research and development in capitalist societies, however, is a precarious venture: while non-exclusive knowledge needs to be transformed into a commodity to enable capitalist accumulation, which per se is already a difficult endeavor, a prosperous economy does require a free flow of information to promote the development and diffusion of innovations. Traditionally, in the field of agriculture, R&D was mainly funded by the public. As we will see later, this is especially true for the U.S. whereas in coordinated economies such as Germany, private companies always played a bigger role in maintaining agricultural R&D activities.

Over the past decades however, the structure of public funding changed dramatically. Significant works of agricultural economists demonstrate a decline in the rate of growth on public spending for agricultural R&D and the shift in the levels of private investment in food and agricultural research compared to public investment [2, 14]. Still, the connection between public spending on R&D and the intended output, such as increased agricultural productivity, is far from clear. Nevertheless, in the current political discourse on national technology strategies only two predictors seem to matter in the evaluation of innovation: R&D expenditures of national governments or private companies and the amount of patents granted in a certain technological field [30].

In this article, we develop a more comprehensive view on innovation in the field of agriculture. Our theoretical perspective refers to the Systems of Innovation Approach, which traces back to the ground-breaking work of Nelson and Winter [32] who offered a new perspective on innovation and technological change. Their work "The Evolutionary Theory of Economic Change" is a fundamental critic on the classic economic model. While the classical model relies on diminishing returns resulting in an equilibrium with each firm making zero profits, Nelson and Winter argue that effective firms show increasing returns to scale, which arise from different types of dynamic behavior as learning by doing [4, 10]. Based on this perspective, the analytical focus shifts from the market mechanism and its potential failures to the firm itself in its interaction with its institutional environment. This also implies that technology development is not characterized by the inevitable unfolding of the most effective type of technology but that technology development is deeply shaped by path dependency rooted in the national innovation system. We do not claim that politics are superior to economy or technology. In fact, our core theoretical argument is that there is a co-evolution of economic institutions and technology development, which unfolds strong dynamics of path dependency [38]. An influx of literature, which examines this connection in a general way, already exists [22, 23, 29, 31]. However, little work has been done in order to get a better understanding of the influences of national innovation systems on the innovation potential in the field of agriculture [53]. Our study tries to advance this debate by focusing on the coordination of agricultural research in two national systems of innovation, the United States and Germany. We argue that, currently, two modes of coordination can be found and that these modes foster the development of different technology types: coordination by the market and coordination by networks.

2 The Difficulty of Organizing Knowledge Production

As we laid out in the introduction, we assume a fundamental incompatibility of knowledge and capitalist accumulation. To overcome this contradiction, institutions that organize the production and provision of knowledge within an economy are necessary. These institutions are the product of a strongly path-dependent process; they emerge from the co-evolution of technology development and institutional adaptation.

2.1 The Contradictions of Knowledge and Capitalist Accumulation

We suggest that there are four aspects of incompatibility between knowledge and its capitalist accumulation [8]. The first two aspects are prominently discussed in neoclassical theory of public goods: knowledge goods tend to be non-excludable and non-rival. To transform non-excludable, non-rival goods into excludable goods, private companies are using two dominant strategies [6]. First, they transform knowledge into a material good that is difficult to reproduce. The combination of software and hardware or the application of hybrid systems in the seed industry is illustrative of this process. Second, strong intellectual property policies, such as patents, enable knowledge protection. To effectively transform knowledge into a commodity, however, a twofold enforcement of intellectual property rights is necessary. Principally, the state needs to provide a mechanism for assigning and enforcing intellectual property rights. The holder of the patent rights must then have the resources necessary to discover when rights are violated and to take legal action against this violation. Therefore, companies usually need to allot considerable resources, such as the establishment of legal departments, to protect their formal intellectual property rights [36, 46]. The third aspect of the incompatibly between knowledge and capitalist accumulation emerges by virtue of the uncertainty in the research process [3]. This uncertainty, however, does not only apply to the research process itself but also to the commercialization of a potential product. The entrepreneur always runs the risk that a competitor is faster in granting a patent, which would temporally result in a market monopoly of the competitor. Therefore, a rational manager (based on cost-benefit analysis) lacks the incentive to invest in research and development. The problem of underinvestment increases the more basic research is involved. Kenneth Arrow [3] concludes that the state should fund research and deliver knowledge as a public good. The fourth aspect of knowledge, which hinders its capitalist exploitation, is somehow different since it cross-cuts the other three aspects. Knowledge can appear in different forms — it can either be explicit or implicit. Implicit knowledge may also be described as tacit, which means that this knowledge cannot be written down and is bound to a person and a certain context. Our argument is that the kind of knowledge predominantly involved in the production of knowledge goods not only influences the excludability of the goods but also the degree of standardization. Since explicit knowledge is knowledge detached from the context of its formation it becomes more or less universally applicable [17]. This means that disentangling the knowledge from a particular context and skills of individual workers makes it more likely that the knowledge good can be standardized. On the production side, standardization leads to savings (economies of scale) because, through explication, the production process and the worker become more manageable and efficient [9]. On the consumption side, standardization implies the expansion of the potential market [43].

In the last paragraph we described four aspects of knowledge that one must overcome to enable capitalist exploitation. Then again, by resorting to the necessary countermeasures, a contradictory dynamic might emerge. To avoid the problem of non-excludability, a state could enforce strong intellectual property rights but it then runs the risk of inducing severe market concentration. To counter the problem of risk aversion of private firms, the state could provide research and development as a public good but this might hinder private investment in this field. To prevent the danger of knowledge spread, a company could keep its production mainly tacit and context-dependent. However, this implies that the potential markets are limited and the firms remain depended on their knowledge workers. The institutional reactions to this dilemma differ highly among states. Different national institutional arrangements offer firms, universities, and other research institutions different sets of opportunities. Or as Hall and Soskice put it: "*there are important respects in which strategy follows structure*" [22, p. 15]. Hence, an in-depth analysis of national institutional arrangements provides the framework for a better understanding of the innovation potential of a nation. As mentioned, there is already a vast body of literature in which these questions are examined in a general way [22, 23, 29, 31], but little work has been done in order to get a better understanding of the influences of national innovation systems in the field of agriculture [53]. In the next section, we suggest two different modes of governance of agricultural R&D within a state: coordination by the market and coordination by networks. Although we make use of the influential framework by Williamson [57] and its further developments [21, 39], we do not think that the coordination types 'markets' and 'networks' exhausts all relevant variation which can be found empirically. In particular, the overuse of 'market' as an analytical concept to describe coordination processes entails a bundle of problematic aspects and fuzziness [7]. Despite these analytical problems, we argue that both analytical concepts market and network help us to develop a better understanding of the interdependencies of national innovation systems and technology development. We study differences in both coordination types with the following characteristics: the design of the intellectual property regime, the division of labor between private and public institutions, the type of intercompany relations, and the dominate type of innovation (mostly based on implicit or explicit knowledge).

2.2 The Coordination of R&D by the Market

Political economy scholars Hall and Soskice [22] argue that in liberal economies, such as the United States, the dominant mode of coordination is the market. This can be seen in several areas: To enable a free flow of labor, the labor market is less regulated than in coordinated economies. Also, the financial system is strongly built on the market mechanism. While banks are more or less outsourced financial departments of companies in coordinated economies, in liberal economies companies must acquire capital on the financial market. Another defining criterion of liberal economies is the competition between companies. This is incorporated by institutional mechanisms like anti-trust policies that limit opportunities for inter-company collaborations and technology diffusion. There are also limited opportunities for private companies to act as a collective when negotiating with government agencies [51]. Coordination of innovation by the market is essentially rooted on the opportunity to transform knowledge into a private (excludable and rival) good. For this transformation, institutions that foster the commodification of knowledge are necessary. Two mechanisms are especially efficient in this regard: enforcement of strong and comprehensive intellectual property rights for applied as well as for basic research and a public funding system that is driven by economic criteria such as royalties of patents or the number of university spin offs.

2.3 The Coordination of R&D by Networks

The second type of R&D coordination follows a very different pattern: the coordination by a network of collaborating firms and state actors. This type of coordination is in line with the general institutional architecture of coordinated economies. Whereas in liberal economies the institutional framework fosters the coordination of innovation by the market, the institutional framework of coordinated economies allows collective organization and bargaining of interest groups, like firms, within one industry [22]. The finance system in coordinated economies is based on banks and concentrated ownership, which allows a long-term horizon in financing of companies [52, 58]. While in liberal economies the competition of companies is a crucial feature, in coordinated economies the instructional framework allows technology exchange and cooperative standard stetting. The analytical concept of networks, as developed by economic sociologists, can help us understand the specific dynamic of technology development in coordinated economies. Williamson [57] stated that the amount of transaction costs determines whether markets or hierarchies are the ideal type of coordination. The economic sociologist Granovetter [21] criticized that neither markets nor hierarchies exist in the real world. This theoretical conception implies a radical empirical perspective on economic action and coordination. Based on the argument that the dichotomy of market and hierarchy does not reflect the reality, sociologist Powell [39] suggested a new type of organization: the network. He stated that next to hierarchies, networks are another powerful mechanism to (at least partially) overcome the threats of opportunism and bounded rationality. With his analytical concept of networks, Powell [39, 40] primarily addressed the new emerging patterns of economic organization, which arise from new technologies and new forms of communication. Despite Powell's different empirical context, we use his analytic insights on networks and apply it to the coordination of R&D in coordinated economies.

2.4 Observable Implications

In order to examine our theoretical argumentation, we will examine the case of the seed market. On the one hand, we will employ a historical perspective to outline the institutional process. On the other hand, we make some basic assumptions that allow us to generate two observable implications for the seed market case.

First, the coordination of R&D by the market privileges innovations that engender highly commercially beneficial production systems. R&D investments have to be amortized by trading and the possibilities for collaborations between companies are very limited. The resulting highly commercially beneficial products have two main characteristics: they contain privately appropriable knowledge and their production shows extremely high returns to scale. We are aware that both criteria are approximately true for every industrial production system. Our point here, however, is that the magnitude of both characteristics is much stronger in liberal production systems. Second, the coordination of R&D by the market encourages radical innovations. In reference to Hall and Soskice [22], we define radical innovations as innovations, which entail substantial shifts in product lines, the development of entirely new goods, or major changes to the production process. Based on this definition, we assume that radical innovations have a higher potential to result in highly beneficial products because they help firms to (at least temporarily) establish a monopoly. While the coordination of R&D by the market fosters fast technological progress and enhances the development of technologies that enable one single company to dominate the market, the coordination of R&D by networks has a stabilizing effect and privileges technologies that maintain the collaboration of existing companies in the network and prevents firms outside the network from participating. Therefore, incremental innovations, which are more reliant on implicit knowledge, are dominant. In the seed sector, the crop type determines whether a product is highly commercially beneficial. For being a highly commercially beneficial crop, three criteria must be met: the possibility to exclude non-paying users (e.g. through hybridization or utility patents), large markets, and finally an innovation which results in a new product that has superior agronomic traits. For the sake of global comparability, we choose two crops, maize and wheat, in the empirical part. We regard maize as a highly commercially beneficial crop. Maize has a natural copy protection through hybridization; thus, the infused knowledge becomes privately appropriable. Moreover, maize is predominately used as feed or energy crop. Therefore, the demand is strongly standardized. Finally, through biotechnology it was possible to create transgenic maize varieties, which bring paramount advantages (such as labor or pesticide savings) to the farmer. Conversely, wheat is an open pollinating cereal, which allows farmers to save their seed. At variance with that for maize, the demand for wheat is strongly diversified. One reason for this is agronomic: especially the winter varieties, wheat demands a stronger customization to different soils than maize varieties. This diminishes the market size for wheat varieties. Another reason is that the demand of bakeries and noodle producers is strongly diversified because they are in need of different characteristics of wheat. Until now, contrary to maize, transgenic wheat varieties did not bring neither agronomic nor economic advantages for the farmers.

Second, next to high profits highly commercially beneficial products show another characteristic: extremely high returns to scale in production and therefore concentration tendencies. In the seed sector, especially the production of transgenic seed shows high returns to scale because the costs for the development and the market approval for transgenic seeds are extremely high compared to the costs of multiplying the seed. However, not only the seed production itself shows high returns to scale but also the fact that genes which carry the desired feature (e.g. herbicide resistance) can be infused in multiple varieties. In accordance with most of the text books on competition theory, we assume that high returns to scale results in concentration tendencies [3, 11, 44].

H1: The innovation activity in highly commercially beneficial crops is higher if the market coordinates R&D.

H2: The coordination of R&D by the market leads to a higher level of market concentration than the coordination by networks.

3 Data and Methods

To validate our theoretical argument, we choose a mixed method approach. We combine a qualitative-historical perspective on national institutions with qualitative expert interviews, and descriptive quantitative data analysis. As we have laid out in the previous paragraphs, we assume that technology development is a co-evolutionary, strongly path-dependent process, which is shaped by the national institutional arrangements as well as from efficient firms which are able to adapt to theses frameworks. The historic perspective helps us to identify the path dependencies of the respective national innovation system by using the example of the seed sector. However, our analyses of the development and the dynamic in the seed industry in the U.S. and Germany are not only based on the works of historians but also on qualitative interviews with experts.[1]

Moreover, to examine our first observable implication, we use data from the Food and Agricultural Organization of the United Nations (FAO STAT) that maintain statistics on annual average crop yields and crop acreage by country. We use yield increase as an indicator of scientific research and innovation activity for the respective crop. Of course, yields are caused by more factors than improved seed, such as input factors (e.g. fertilizers, herbicides, and pesticides) and improved cultivation methods. However, the quality of seed is regarded to be a core factor of agronomic performance [12]. Although our primary comparison is between the U.S. and Germany, we include data on a few other countries to highlight the differences between the U.S. and Germany. To examine our second observable implication, we compare the levels of concentration in the agricultural seed sector in the U.S. and Germany using the Herfindahl Hirschman Index (HHI). The HHI is the most commonly used measure of concentration also applied by the U.S. anti-trust authority. It is an absolute measure of concentration, which reports of the sum of the squared market shares. For an HHI between 1000 and 1800, the market is considered to be concentrated. An HHI score above 1800 indicates substantial concentration [45]. Schenkelaars et al. [45] have calculated the HHI score for the U.S. and we use that score. Nevertheless, the main limitation for public research in agricultural markets is, generally, the availability of data [13]. Thus, there is no public available study on concentration of the German seed sector, probably because firm data is limited due to strategic reasons of the companies [13]. We try to overcome this problem by the use of two new datasets: First, we calculate the HHI using the database of the German Maize Committee to estimate the market concentration in maize seed. The German Maize Committee tests varieties and publishes the results to assist farmers in purchasing decisions. Additionally, we provide new data on the concentration tendencies in other seed varieties using seed approvals [Sortenzulassung] of the German Federal Plant Varieties Office [Bundessortenamt] for the time period 1990–2010.[2]

[1] We interviewed approximately 60 persons who were located in the U.S. and in Germany. As experts we regard mangers of breeding firms/agrochemical companies, breeders, scientists at universities in the field of Biology, members of governmental authorities, and farmers.

[2] Ideally, one would have one common data source. However, given the before mentioned data limitation, there is no data source that includes market data both from Germany and the U.S. A further limitation is that the seed approval procedure is not harmonized within the European Union. Hence, seed sorts approved in other European countries may also be traded in the German market. However, according to our interview partners, a German approval functions as a quality signal for the German market and therefore, bias should be small. This is not the case for maize, and therefore, we rely on the data from the Maize Committee for this crop. In general, reliable inferential data analysis would demand panel data on the firm level, which is not available for any country [13].

4 Results

4.1 The U.S. Case: Coordination of Agricultural R&D by the Market

The statement that R&D in the U.S. is coordinated by the market may provoke opposition, which would be very justified. Indeed, during the cold war area, huge parts of the U.S. public research and development activities were not driven by the market but by 'great visions' such as landing humans on the moon or the creation on a superior military complex [15, 41]. This context also applies to the enormous increase in U.S. crop yields during the 1940 and 1950s as well as the Green Revolution.

During the Cold War, public development of seed was not only understood as a means of securing domestic food supply but also as a weapon. Within a very brief period, the U.S. investments in wheat breeding made many countries (including the Soviet Union) dependent on U.S. wheat exports [1, 37, 59]. However, the research university in general and the land-grant system in particular underwent dramatic changes starting late 1970s [26, 49, 50]. The fundamental restructuring of the university system was caused by the political détente towards the end of the Cold War as well as the growing economic weakness of the United States. The economic shortcoming of the U.S. was related to the global diffusion of U.S. technology especially to its competitors, Japan and Germany [33, 47]. This change resulted in a university system that operates on the same logic as private companies: the market mechanism. This 'new type' of coordinating university activities is in line with the architecture of the institutional framework of liberal economies. Until the 1970s there has been almost no protection for intellectual property in the plant breeding sector.[3]

The Plant Variety Protection Act from 1970 implemented the first intellectual property certificates for crops. These protection rights were still rather weak compared to other countries as Germany. These weak intellectual property certificates were complimentary to the coordination of agricultural R&D by the state. Only some small areas of R&D were exempted from state coordination, such as the development of seeds in crops for which hybridization is possible, e.g. corn. However, the slow withdrawal of the state from coordinating agricultural R&D after the end of the Cold War as well as increased costs for seed development through biotechnology made a stronger intellectual property protection necessary. No other legislation was as groundbreaking as the Bayh-Dole Act of 1980. The Bayh-Dole Act renegotiated the question how to treat intellectual property that arises from federal public funding. Before Bayh-Dole, all research results and inventions from universities or public

(Footnote 2 continued)
Hence, we provide descriptive and preliminary results. Nevertheless, we carefully selected data available and believe we are able to show general tendencies.

[3] The Plant Patent Act from 1930 only applies to a-sexual reproduced plants, which means that basically all crops are excluded.

research institutions were considered to be public goods. Contrary to this open access policy, Bayh-Dole permitted universities to grant patents — even on basic research. In the seed sector the judgment 'ex parte Hibberd' of 1989 in accordance with the Bayh-Dole Act allowed firms and universities to grant utility patents on seeds or even single genes. The enforcement of stronger intellectual property rights not only allowed better appropriability of innovations, it also drove the commodification of innovation. Thus, intellectual property rights enabled firms to include innovation in a formal decision making process which is based on a monetary cost-benefit analysis as well as to enter into contracts with other companies on the use of certain technologies, such as cross licensing or mergers, and acquisitions. At the level of universities the coordination of R&D by the market had similar implications. As already mentioned, in the area of Cold War, research at federal funded universities was driven by 'great visions' rather than by economic profitability. In the seed sector therefore, a division of labor between the private and the public sector evolved [54]. The public sector was responsible for basic research and the development of seed for minor and less beneficial crops such as wheat or barley. The private sector focused on breeding activities among commercially profitable crops such as corn or cotton. In the 1980s, the restructuring of the university system blurred these boundaries. The subjection of university research to the market logic changed the criteria for academic excellence towards marketability [50]. The transformation manifested itself in the move of university researchers towards more commercially relevant crops [54], a decline in number of publicly employed plant breeders [5], and an increase in the proportion of research results in the public sector protected by patents [42]. During that same period, also the broader agricultural research and development structure changed. Agrochemical companies such as Monsanto or Dow acquired the majority of the medium sized seed companies and invested heavily in biotechnological research and development [24, 45]. Moreover, the number and scope of university-industry research collaborations expanded [19].

4.2 The German Case: Coordination of Agricultural R&D by Networks

As aforementioned, coordinated economies provide an institutional structure that creates space for collective organization and bargaining of interest groups. In the seed sector, two factors are of particular importance for understanding how the institutional structure affects the innovation process: the design of intellectual property law and the regulation of the market by state actors. Plant Variety Protection [Sortenschutz] is an industry-specific form of intellectual property protection that supports the cooperative structure of the sector. By the end of the 1920s, the Association of German Plant Breeders was able to enforce the implementation of intellectual property rights in the seed sector. This was very early compared to other countries [55]. The early implementation of intellectual property rights fostered the accountability of innovations, enabling German breeders to establish their brand names in the market. These early versions of plat variety protection from the 1920s were updated and, in 1953, the Plant Variety Protection [Sortenschutzgesetz] which allows breeders to access all existing varieties when doing research and covers the entire plant genome, was enacted [28]. Compared to the relevant U.S. legislations, however, farmers' rights under the German law were less comprehensive [27]. The German intellectual property right is not only complementary to the predominate type of innovation, but it also promotes cooperation amongst companies in the German seed sector. Another important aspect is that the German seed market is subjected to strong governmental regulation and artificial market interventions. The Federal Plant Variety Office not only grants plant variety protection but also makes decisions on market approval for the respective variety. The Federal Plant Variety Offices explicit objective is not the provisioning of variety diversity, but the adjustment of the market. The office has determined that it should be easy for farmers to decide upon the variety and that only high yielding varieties should be offered. Historically, this role arose during a period when increasing domestic agricultural production was the primary goal [16].

These characteristics of the German institutional framework foster the collaboration between medium-sized, mostly family-run breeding firms. The network does not only include mangers and breeders of collaborating firms, however, but also university researchers and members of governmental authorities (e.g. the Plant Variety Protection Office or the German Ministry of Agriculture). This network is strongly based on long-term personal relationships, which have been handed down over generations. Scientific societies, as well as industry networks in coordinated economies, are traditionally more oriented on the ideal of medieval guilds that they get privileges from the state as reward for their social service in education, standard setting, and the preservation of quality for goods and services. While guilds are deeply committed to a professional ethic and, therefore, have a high obligation to public goods, modern businesses clubs, or lobby groups try to enforce their group interests in the political arena. A second factor, which contributes to the stability of the majority of networks in coordinated economies, is the relatively small amount of members [34], which increases the accountability of individual action and, thereby, reduces opportunism. However, this stability also has a downside. It prevents external actors from accessing the industry. Therefore, the coordination of R&D by a close network hinders (or at least delays) the application of new (or just different) technologies. The closeness of the network in the seed sector leads to a paradoxical situation: ecological breeders encounter the same barriers as biotechnological breeders.

4.3 Innovation Activity in Different Crops

As we laid out in Sect. 2.4, we assume that some crops are more commercially beneficial than others. For the sake of global comparability we chose two crops, maize (highly commercially beneficial) and wheat (less commercially beneficial) to

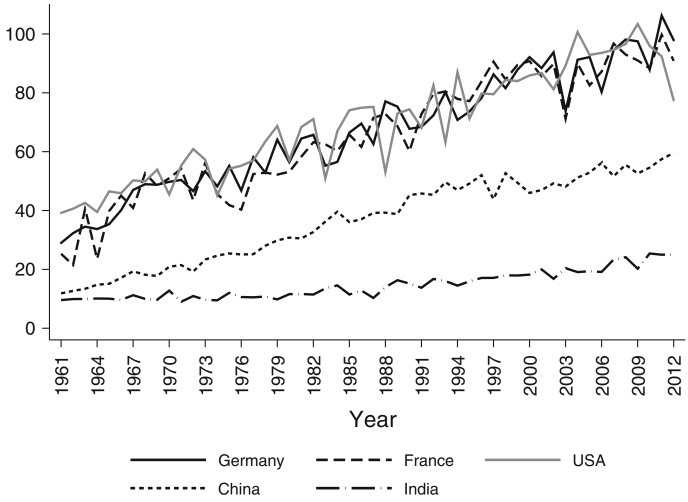


Fig. 1 Development of yield per hectare in maize (in 1000 hg), *source* FAO STAT, own calculation

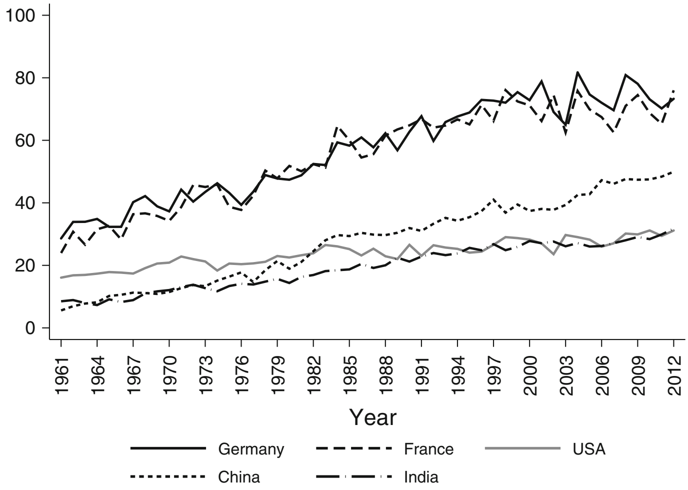


Fig. 2 Development of yield per hectare in wheat (in 1000 hg), *source* FAO STAT, own calculation

examine our hypothesis. In Figs. 1 and 2 we see the average development of yield per hectare hectogram in maize and wheat from 1961–2013.

It is important to look at the respective growth rate instead of comparing the absolute differences. Hence, the differences at the beginning can be explained by durable conditions such as the quality of soil or climate; however, the annual growth rates can basically be explained by an improved quality of seed. The charts show remarkable differences in the developments of yields in different crops. In China and India, the growth rate of wheat is much higher than in maize. This can be explained by the demand for the important food crop wheat and the high level of public funding in agricultural R&D in both countries [35, 48]. Alternatively, the demand of meat and, for that reason, for the feed crop maize rose only in the last decade and simply in China. In Germany and France, seed development mainly takes place in collaborating medium-sized breeding companies. In these countries, the demand for feed and food crops is equal and the annual rate of increase in maize and wheat is also similar.

However, in the U.S., the rate of yield increase in maize is much higher than that in wheat. We already mentioned that there are more reasons for the low wheat yields next to the quality of seeds; however, the U.S. wheat yields are remarkable low in an international comparison. They are much behind the Chinese wheat yield, whereas the Indian wheat yields are almost on the same level. For crops that are more commercially beneficial, such as soy or rice, the rate of yield increase is much higher in the US than in India [18]. The low U.S. wheat yields are even more surprising when we note that the almost the whole world was dependent on U.S. wheat exports between the 1940 and 1960s. From 1937 up to 1964, U.S. foreign trade surplus in wheat rose from 1.1 to 40.7 million tons, which was deemed to be four-fifth of the total world trade [1]. This incredible production increase was possible due to the immense public funding in this time [37, 59].

4.4 Market Concentration

While the U.S., as well as the global seed market, underwent a dramatic process of market concentration [24, 45], we could not find these tendencies in the German seed market. Schenkelaars [45, p. 18] shows that in 1985, the nine biggest companies in the seed market had a market share of 12.7%. This share rose to 16.6% in 1996. In 2009, the three biggest companies (Monsanto, DuPont, and Syngenta) had a share of 34%. In Fig. 3, we see that the Herfindahl Hirschman Index in the market for maize seed is much higher in the U.S. market. When we look at the firms that are active in the market, we see that the big agrochemical companies (such as Du Point Pioneer or Bayer) participate in the German seed market too but they do not have a dominating position. In Fig. 4, we observe that the HHI in the wheat and barley market is below 1000 (*dashed line*) during almost the entire investigation period. The concentration in the rapeseed market was at 2500 in the early 1990s. However, in the last 15 years it decreased dramatically with an HHI now being at about 1500 points. In comparison to wheat and barley, rapeseed is hybrid and therefore, commercially more beneficial.

As discussed in Sect. 2.4, we regard the higher market concentration in the U.S. market as a result of the organization of R&D. In the institutional context of the United States, companies naturally invest in technologies wherein private appropriation is possible and large markets are existent. Transgenic seed encounters both requirements, and herbicide tolerance and insecticide resistance are traits that make the seed superior in the context of the highly industrialized U.S. agriculture. In the

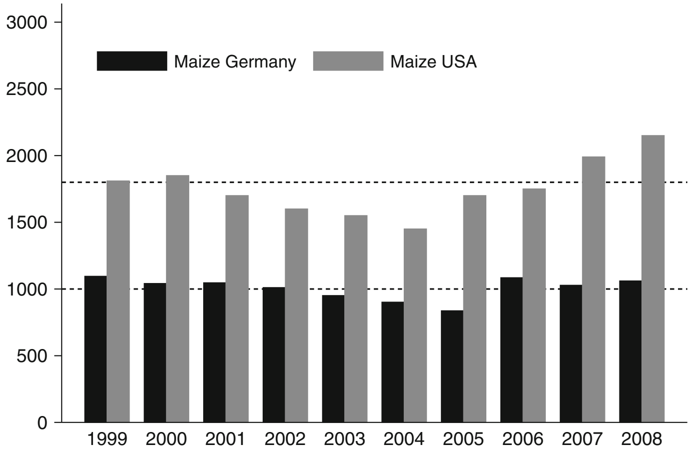


Fig. 3 Herfindahl Hirschman Index in USA and Germany, *source for Germany* data provided by German Maize Committee 2013, own calculation, source for the U.S.: Schenkelaars et al. [45, p. 43]

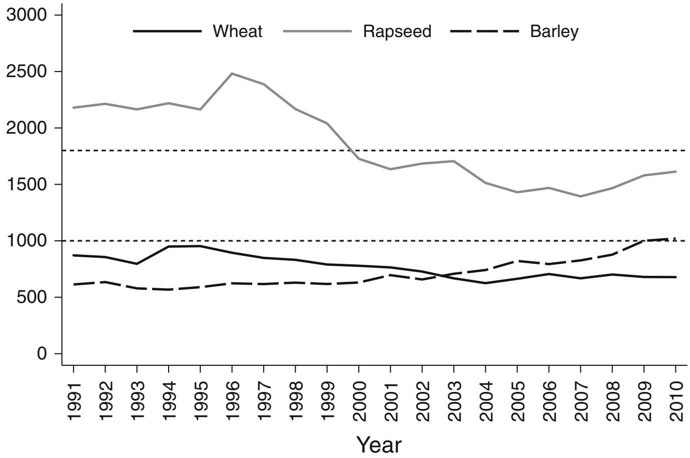


Fig. 4 Herfindahl Hirschman Index for non-maize crops in Germany, *source* data provided by the German Federal Plant Varieties Office, own calculation

production phase, transgenic seeds show extremely high increasing returns to scale because the costs for development and market approval are extremely high when compared to the costs of multiplying the seed. Then again in Germany, R&D in the seed sector is coordinated by a network. This network is based on stable personal relations, which hinder the diffusion of knowledge to competitors. The innovation process in the German seed sector is predominately based on implicit knowledge.

This implies that the innovations are not radical but incremental. The increasing returns to scale for each innovation step are much smaller than in the U.S. case and therefore, the concentration tendencies are much lower.

5 Conclusion

We have laid out the co-evolutionary process of technology development and institutions in the previous section. These theoretical insights may help us to develop a better understanding of the ban on transgenic seed in Germany. Contrary to the flat narrative that no transgenic seed is cultivated in Europe because of (irrational) consumers, who were able to organize efficient protest and boycott, we suggest that the ban of transgenic seed also has structural reasons. The German innovation system is based on implicit/ incremental innovations and collaboration. The adaptation of biotechnological methods would question this cooperative agreement of medium-sized breeding firms. Second, the technology historian Wieland [56] shows that the delayed and reluctant reception of biotechnological methods in the German industry can be explained by a path-dependent process, which was pre-structured by the chemical industry. The German industrys hostility towards biotechnology contradicts the early promotion of this field by the state. Graff, et.al. [20] lay out that U.S. American companies have a comparative advantage in biotechnological innovations, while German firms have a cooperative advantage in chemical innovations. Therefore, protesting consumers, farmers unions, and the German agrochemical companies become ‘strange bedfellows’ in fighting against transgenic plants. In reference to Hall and Soskice [22], we want to emphasize that there is not one best institutional arrangement, which leads to economic success in a late capitalist economies but there are some best (institutional) answers, which developed co-evolutionary to the requirements of technology and social forces. Hence, the structure of public R&D funding as well as political interventions, which aim to change the institutional frameworks as e.g. the global homogenization of intellectual property rights, have to orientate itself towards the national system of innovation rather than on a global agenda.

We would like to conclude with an illustrative case on how the very different organization of agricultural R&D in Germany and Canada responded to the same technical problem. Rapeseed was not suitable for human consumption until the 1970s. This changed when a German rapeseed breeder accidentally found an erucic-acid free rapeseed mutant. At that time, all five German rapeseed breeding companies cooperated with each other and secured government funding for further development. Each company, however, performed the last steps in developing a marketable variety on its own. In 1981, the first erucic-acid free variety was released. Cooperation among the companies was not only useful for pooling R&D resources; it was also necessary to reduce cross-pollination, which is especially high in rapeseed. It would have been unlikely that one breeder alone had developed a new variety. In Canada, an economy that is more consistent with the liberal type, the development of rapeseed varieties suitable for human consumption followed a very different path. The erucic-acid free varieties (Canola) were exclusively invented by public breeding programs in the 1990s [25]. Contrary to the German case, the Canadian government supported the application of transgenic methods to improve the rapeseed varieties. During the 1990s, transnational agrochemical companies purchased these public breeding programs. While Monsanto and Bayer dominate the Canadian rapeseed (Canola) market, the German market was still predominantly lined by the same medium-sized companies.[4]

References

1. Abel, W.: Agrarpolitik. Vandenhoeck and Ruprecht, Göttingen (1967)
2. Alston, J.M.: Persistence pays. U.S. agricultural productivity growth and the benefits from public R&D spending. Springer, New York (2010)
3. Arrow, K.J.: Economic welfare and the allocation of resources for invention. In: N.B.C. for Economic Research (ed.) The rate and direction of inventive activity: economic and social factors — a conference of the Universities-National Bureau Committee for Economic Research and the Committee on economic Growth of the Social Science Research Council, pp. 609–626. University Press, Princeton (1962)
4. Arthur, W.B.: Increasing Returns and Path Dependence in the Economy. University Press, Ann Arbor (1994)
5. Bliss, F.: Education and preperation of plant breeders for careers in global crop improvement. Crop Sci. **47**, 250–261 (2007)
6. Böschen, S., Brandl, B., Gill, B., Spranger, P., Schneider, M.: Innovationsförderung durch geistiges eigentum? passungsprobleme zwischen unternehmerischen wissensinvestitionen und den schutzmöglichkeiten durch patente. In: Grande, E., Jansen, D., Jarren, O., Rip, A., Schimank, U., Weingart, P. (eds.) Neue Governance der Wissenschaften. Transcript, Bielefeld (2013)
7. Boyer, R.: The variety and unequal performace of really existing markets: farwell to doctor pangloss? In: Hollingsworth, J.R. (ed.) Contemporary Capitalism. The Embeddedness of Institutions, pp. 55–93. University Press, Cambridge (1997)
8. Brandl, B., Paula, K., Gill, B.: Spielarten des wissenskapitalismus. die kommodifizierung von saatgut in den usa und in deutschland. Leviathan **4**(42), 539–572 (2014)
9. Braverman, H.: Labor and Monopoly Capital. The Degradation of Work in the 20th Century. Monthly Review Press, New York (1974)
10. David, P.A.: The dynamo and the computer: an historical perspective on the modern productivity paradox. Am. Econ. Rev. **80**(2), 355–361 (1990)
11. Demsetz, H.: The systems of belief about monopoly. In: Goldschmid, H., Mann, M., Weston, F. (eds.) Industrial Concentration: The New Learning, pp. 164–183. Boston, Brown, Little (1974)
12. Fernandez-Cornejo, J.: The seed industry in U.S. agriculture: an exploration of data and information on crop seed markets, regulation, industry structure, and research and development. Technical report, United States Department of Agriculture, Economic Research Service. Agriculture Information Bulletin No. 33671 (2004)
13. Fernandez-Cornejo, J., Just, R.E.: Researchability of modern agricultural input markets and growing concentration. Am. J. Agric. Econ. **89**(5), 1269–1275 (2007)

[4] The five companies are: NPZ Lemke, DSV, W.v. Borries-Eckendorf, Raps GbR, and KWS Saat. In 2011 Bayer acquired Raps GbR. Other companies perceived this acquisition very negatively. Until now however, this acquisition did not change the market structure. Thus, the market share of Raps GbR was very small, anyway.

14. Fuglie, K., Heisey, P., King, J., Pray, C.E., Schimmelpfennig, D.: The contribution of private industry to agricultural innovation. Sci. **338**(6110), 1031–1032 (2012)
15. Galison, P., Hevly, B.W.: Big Science: The Growth of Large-Scale Research. University Press, Stanford (1992)
16. Gill, B., Brandl, B.: Rechtsschutz von Pflanzenzüchtungen. Eine kritische Bestandaufnahme des Sorten-, Patent- und Saatgutrechts, chap. Legitimität von Sortenschutz und Sortenzulassung aus soziologischer Sicht, pp. 163–186. Mohr Siebeck, Tübingen (2014)
17. Gill, B., Brandl, B., Böschen, S., Schneider, M.: Autorisierung. eine wissenschafts — und wirtschaftssoziologische perspektive auf geistiges eigentum. Berl. J. für Soziol. **22**(3), 407–440 (2012)
18. Glenna, L., Brandl, B., Jones, C.: International political economy of agricultural research and development. In: Bennano, A., Bush, L. et al., (eds.) Handbook of International Political Economy of Agriculture and Food. Edward Elgar, Cheltenham
19. Glenna, L.L., Lacy, W.B., Welsh, R., Biscotti, D.: University administrators, agricultural biotechnology, and academic capitalism: defining the public good to promote university industry relationships. Sociol. Q. **48**(1), 141–163 (2007)
20. Graff, G.D., Hochman, G., Zilberman, D.: The political economy of agricultural biotechnology policies. AgBioForum **12**(1), 34–36 (2009)
21. Granovetter, M.: Economic action and social structure: the problem of embeddedness. Am. J. Sociol. **91**(3), 481–510 (1985)
22. Hall, P.A., Soskice, D.W.: Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. University Press, Oxford (2001)
23. Hollingsworth, J.R.: Contemporary Capitalism; The Embeddedness of Institutions. University Press, Cambridge (1997)
24. Kalaitzandonakes, N., Magnier, A., Miller, D.: A worrisome crop? is there a market power in the u.s. seed industry? Regul. **20**, 20–26 (2011)
25. Kinchy, A.J.: Seeds, Science, and Struggle: The Global Politics of Transgenic Crops. The MIT Press, Cambridge (2012)
26. Kleinman, D.L.: Politics on the Endless Frontier: Postwar Research Policy in the United States. University Press, Durham (1995)
27. Kloppenburg, J.R.: First the Seed. The Political Economy of Plant Biotchnology. University Press, Cambridge (1988)
28. Leßmann, H., Würtenberger, G.: Deutsches Und Europäisches Sortenschutzrecht: Handbuch, 2nd edn. Nomos, Baden-Baden (2009)
29. Lundvall, B.A.: National Systems of Innovation: Toward a Theory of Innovation and Interactive Learning. Anthem Press, London (2010)
30. Mazzucato, M.: The Entrepreneurial State: Debunking Public Versus Private Sector Myths. Anthem Press, London (2013)
31. Nelson, R.: National Innovation Systems: A Comparative Analysis. University Press, Oxford (1993)
32. Nelson, R.R., Winter, S.G.: An Evolutionary Theory of Economic Change. University Press, Harvard (1982)
33. Nelson, R.R., Wright, G.: The erosion of U.S. technological leadership as a factor in postwar economic convergence. In: Baumol, W.J., Nelson, R.R., Wolff, E.N. (eds.) Convergence of Productivity: Cross-National Studies and Historical Evidence, pp. 129–163. University Press, Oxford (1994)
34. Olson, M.: The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge (1971)
35. Pal, S., Byerlee, D.: India: the funding and organization of agricultural R&D–evolution and emerging policy issues. In: Pardey, P.G., Alston, J.M., Piggott, R. (eds.) Agricultural R and D in the Developing World: Too Little, Too Late?, pp. 155–193. International Food Policy Research Institut, Washington D.C. (2006)
36. Pechlaner, G.: Corporate Crops: Biotechnology, Agriculture, and the Struggle for Control. University of Texas Press, Austin (2012)

37. Perkins, J.H.: Geopolitics and the Green Revolution: Wheat, Genes, and the Cold War. University Press, Oxford (1997)
38. Pierson, P.: Increasing returns, path dependence, and the study of politics. Am. Polit. Sci. Rev. **2**(94), 251–257 (2000)
39. Powell, W.: Neither market nor hierarchy: network forms of organization. Res. Organ. Behav. **12**, 295–336 (1990)
40. Powell, W., Grodal, S.: Networks of innovators. In: Fagerberg, J., Mowery, D.C., Nelson, R. (eds.) The Oxford Handbook of Innovation. University Press, Oxford (2006)
41. Reynolds, D.: Science, technology, and the cold war. In: Leffler, M.P., Westad, O.A. (eds.) The Cambridge History of the Cold War, vol. 3, pp. 378–399. University Press, Cambridge (2010)
42. Rhoten, D., Powell, W.: Public research universities from land grant to patent grant institutions. In: Rhoten, D., Calhoun, C.J. (eds.) Knowledge Matters.The Public Mission of the Rresearch University, pp. 319–345. University Press, Columbia (2011)
43. Ritzer, G.: The McDonaldization of Society. Pine Forge Press, Thousand Oaks (1993)
44. Rubinfeld, D.: Antitrust policy. In: Smelser, N.J., Baltes, P.B. (eds.) International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Pergamon, Oxford (2001)
45. Schenkelaars, P., Vriend, H., Kalaizandonakes, N.: Drivers of consolidation in the seed industry and its consequences for innovation. Technical report, Schenkelaars Biotechnology Consultancy for COGEM (report CGM2011-11) (2011)
46. Schubert, J., Böschen, S., Gill, B.: Having or doing intellectual property rights? Transgenic seed on the edge between refeudalisation and napsterisation. Eur. J. Sociol. **52**(1), 1–17 (2011)
47. Scotchmer, S.: The political economy of intellectual property treaties. J. Law Econ. Organ. **20**(2), 415–437 (2004)
48. Shenggen, F., Qian, K., Zhang, X.: China: an unfinished reform agenda. In: Pardey, P.G., Alston, J.M., Piggott, R. (eds.) Agricultural R and D in the Developing World: Too Little, Too Late?, pp. 29–63. International Food Policy Research Institut, Washington D.C. (2006)
49. Slaughter, S., Rhoades, G.: The emergence of a competitiveness research and development policy coalition and the commercialization of academic science and technology. Sci. Technol. Hum. Values **21**(3), 303–339 (1996)
50. Slaughter, S., Rhoades, G.: Academic Capitalism and the New Economy: Markets, State, And Higher Education. Johns Hopkins University Press, Baltimore (2004)
51. Soskice, D.W.: Divergent production regimes: coordinated and uncoordinated market economies in the 1980s and 1990s. In: Kitschelt, H., Lange, P., Marks, G., Stephens, J. (eds.) Continuity and Change in Contemporary Capitalism, pp. 101–134. University Press, Cambridge (1999)
52. Streeck, W.: Re-forming Capitalism: Institutional Change in the German Political Economy. University Press, Oxford (2009)
53. Vanloqueren, G., Baret, P.V.: How agricultural research systems shape a technological regime that develops genetic engineering but locks out agroecological innovations. Res. Policy **38**(6), 971–983 (2009)
54. Welsh, R., Glenna, L.: Considering the role of the university in conducting research on agribiotechnologies. Soc. Stud. Sci. **36**(6), 929–942 (2006)
55. Wieland, T.: Wie beherrschen wir den pflanzlichen Organismus besser, ... In: Wissenschaftliche Pflanzenzüchtung in Deutschland, 1889–1945. Deutsches Museum, Munich (2004)
56. Wieland, T.: Pfadabhängigkeit, forschungskultur und die langsame entfaltung der biotechnologie in der bundesrepublik deutschland. In: Fraunholz, U., Hänseroth, T. (eds.) Ungleiche Pfade? Innovationskulturen Im Deutsch-Deutschen Vergleich, pp. 73–98. Waxmann, Münster (2012)
57. Williamson, O.: Markets and Hierarchies. Free Press, New York (1975)
58. Windolf, P., Beyer, J.: Kooperativer kapitalismus. Kölner Zeitschrift für Soziol. Sozialpsychologie **47**(1), 1–36 (1995)
59. Wright, B.D.: Grand missions of agricultural innovation the need for a new generation of policy instruments to respond to the grand challenges. Res. Policy **41**(10), 1716–1728 (2012)

Short and Long Run Armington Elasticities for the Mexican Economy

Enrique R. Casares, Lucía A. Ruiz-Galindo and Horacio Sobarzo

Abstract The Armington elasticity is a key element in models with trade flows, either in International Real Business Cycle (IRBC) models or in computable general equilibrium models. In this paper, Armington elasticities at the aggregate level are estimated for Mexico for the 1993–2013 period. The composite good, formed by domestic and imported goods, is defined by means of an aggregate social accounting matrix for Mexico. This composite good is modeled through of a constant elasticity of substitution function. The relative demand for imports to domestic goods is obtained as a function of their relative prices. The two variables of the model, the logarithm of the relative demand for imports to domestic goods and of their relative prices, are integrated of order one and cointegrated. Therefore, an error correction model is used in order to obtain short and long run elasticities. Thus, short and long run elasticities are 0.534 and 0.719, respectively. The estimated elasticities are consistent with those used in IRBC models, which are relatively small elasticities. Also, long run elasticity is higher than short run elasticity, as presented in the literature.

Keywords Real business cycles models · Computable general equilibrium models · Armington assumption · Unit root · Cointegration · Error correction model

We thank to Elizabeth Viveros-Vergara for her valuable help.

E.R. Casares (✉) · L.A. Ruiz-Galindo
Departamento de Economía, Universidad Autónoma Metropolitana
Unidad Azcapotzalco, Av. San Pablo 180, Col. Reynosa Tamaulipas,
02200 Delegación Azcapotzalco, Ciudad de México, D.F., Mexico
e-mail: ercg@correo.azc.uam.mx

L.A. Ruiz-Galindo
e-mail: laruizg@correo.azc.uam.mx

H. Sobarzo
El Colegio de México, Centro de Estudios Económicos, Camino al Ajusco 20,
Tlalpan, Pedregal de Santa Teresa, 10740 Ciudad de México, D.F., Mexico
e-mail: hsobarzo@colmex.mx

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_6

1 Introduction

In models with trade flows, one of the key elements to understanding the behavior of the trade variables and macroeconomic aggregates is the elasticity of substitution between domestic and imported goods, also known as Armington elasticity, which essentially posits that goods are different depending on the place they are produced, and therefore, rarely are perfect substitutes when their prices fluctuate [2].

Specifically, the importance of Armington elasticities has been evident in two branches of economic modeling, International Real Business Cycles (IRBC) models and Computable General Equilibrium (CGE) models, which have different perspectives on their values, and, in both cases, the arguments are reasonable.

On the one hand, to account for the volatility of the terms of trade and movements in the balance of trade, in IRBC models, the practice is to use relatively small calibrated elasticity values in a range between 0.5 and 2 [7, 8]. For example, [3] used a value of 1.5 for the Armington elasticity of substitution (also, [18]). In addition, to provide empirical support for the use of these small elasticities, [13] estimate the substitution elasticity between domestic and foreign goods with aggregate data, finding a value of 0.9. Similarly, [4] estimate an elasticity of 1.13 with data at a macro level.

On the other hand, to explain the growth of international trade, in CGE models, it is very common to find values above 3 and up to 6, depending on different estimations ([1]). [22], for example, uses values of 3 and 4 for the goods considered more tradable, such as in agriculture.

Thus, there seems to be a significant discrepancy regarding these key values when studying the flows of trade. Some reasons that help explain these discrepancies are related to aspects such as the level of aggregation of goods. Thus, in IRBC models aggregation levels are high, while CGE models usually operate with higher levels of disaggregation.

This seemingly unsolvable contradiction has, in fact, a very coherent explanation. In this regard, [21] provides an argument that reconciles both approaches. The idea is simply that the elasticities evaluated with high frequency data on prices and quantities, such as in IRBC models, capture responses to transitory shocks to productivity or demand. On the other hand, the elasticities estimated with changes in trade policy, as in CGE models, are capturing responses to permanent shocks. Therefore, as the agents react differently to permanent or temporary changes, it is normal, then, that the elasticities would differ.[1]

The objective of this paper is to quantify the degree of substitution between imported and domestic goods in Mexico, caused by changes in the relative prices of these goods; that is, the Armington elasticities are estimated at the aggregate level, as in IRBC models. This estimation is important, not only for being the first of its kind for Mexico, but, also, because it contributes to a better understanding of

[1] Hertel et al. [15] point out some problems with econometric estimation techniques and the mismatch between the data sample, the source of variation in the econometric estimation and the policy experiment to be performed in CGE models.

macroeconomic fluctuations associated with the phenomena of trade liberalization and changes in the terms of trade. The value of this elasticity is key to measuring social welfare impacts associated with these phenomena.

In this paper, we present an aggregated Social Accounting Matrix (SAM) for Mexico with the purpose of defining the so-called composite good, consisting of domestic and imported goods. The quantity of the composite good is defined by a Constant Elasticity of Substitution (CES) function, which describes the preferences of consumers to substitute imported for domestic goods. More specifically, a representative domestic consumer is considered to minimize the expenditure on domestic and imported goods subject to the CES function. The first order conditions for this optimization problem lead to the relative demand of imported goods to domestic ones as a function of their relative prices, which is the first model to be considered for the estimation of the Armington elasticities. Together with this, alternative models are formulated, like the partial adjustment and error correction models, which make sense when one has analyzed the properties of the data incorporated in the model.

In the estimation of the model, we use quarterly data corresponding to the domestic supply and imports with their respective price indices for the period from the first quarter of 1993 to the fourth quarter of 2013. Therefore, it is essential to study the order of integration of the variables and their possible cointegration in order to choose between alternative models so we can make a proper analysis of the short and long run effects. The specification of the final model also depends on the economic and econometric evaluations. The latter is an important feature of this work, because a detailed assessment of the assumptions of the econometric model is carried out. Most of the literature on the subject confine themselves to analyzing only the statistical significance of the elasticities, or simply presenting their values.

Due to the fact that the two variables of the model, the logarithm of the relative demand of imports to domestic goods and of their relative prices, are integrated of order one and cointegrated, the estimations were made using an error correction model. Thus, Armington short and long run elasticities estimated in this paper are 0.534 and 0.719, respectively. These results are related to the articles of [11, 16], where using cointegration techniques allow them to distinguish between short and long run elasticities. In addition, [11] find that long run elasticities of import demand are usually higher than short run elasticities, as presented in this paper. Also, our results show that the substitution elasticities between domestic and imported goods are less than one, as in [13].[2]

The paper is organized as follows. In Sect. 2, an aggregated SAM is shown to define the composite good, and to explain how the time series used in the econometric estimation were constructed. In Sect. 3, the optimization problem is presented, and the relative demand model is obtained. In Sect. 4, the econometric methodology and the different models that can be estimated according to the statistical properties of empirical information are presented. In Sect. 5, once the stationarity analysis is

[2] Crucini and Davis [5] develop a model where the discrepancy between the values of substitution elasticity, the short and long run, is a result of the frictions in the distribution sector.

completed and the order of integration of the variables is determined, the most appropriate model is chosen, estimated and evaluated. The paper ends with conclusions in Sect. 6.

2 Social Accounting Matrix

In this section, the formation of a Social Accounting Matrix (SAM) for Mexico for the year 2003 is described, where all goods are aggregated into only one, which describes in detail how the supply of (and demand) this good is formed. The idea is simply to define the composite good and show where the optimal choice, between domestic and imported goods, is presented. At the same time, we show how the time series used in the econometric estimation are constructed. In addition, the SAM format arrangement tries to act as point of reconciliation between the two approaches alluded in the previous section, a Macro SAM that, eventually, when disaggregated, provides the basis for a CGE model.

The SAM presented here was developed using the information contained in the Mexican Input-Output Matrix for 2003 and in the Mexican National Accounts for the same year (INEGI). The SAM is formed by the productive activities, goods, factors of production, institutions, investment, and the rest of the world accounts (Table 1).

As is well known, a SAM is a square accounting format, where economic activities of the main actors of the economy are recorded in monetary terms for a given period. A row (sales) and a column (purchases) correspond to each agent. The row represents the income, and the column represents expenses, this way reflecting, for each agent, the accounting identity that income is equal to expenditure.

In Table 1, one can observe that the account of productive activities is broken down into two sub-accounts, value added and gross output. The first sub-account corresponds to the net value of the production of goods (column 1), which consists of payments to factors (labor and capital) and tax payments on production. In the second sub-account (column 2), the formation of gross production is described, which is made up of net output and intermediate consumption (row 6, column 2).

The formation of the supply is described in columns 3, 4, 5 and 6. Thus, taxes on products (393) are added to the domestic consumption supply (10,509) to form the internal supply at market prices (10,902). In turn, the supply of imports is recorded in column 5, where purchases from the rest of the world are recorded (2,026). Thus, the two sources of supply, domestic and imported, are shown at market prices. Both are reported in column 6 to form the total supply (12,928) for domestic consumption. As noted, this supply is called composite good, because it brings together the two sources of supply for the country, domestic and imported. This is where the optimal choice between domestic and imported goods arises, modeled through of the Armington assumption. Note that a part of domestic production goes to the rest of the world in the form of exports (row 2, column 4).

Table 1 Social accounting matrix of 2003 (billions of Pesos)

			1	2	3	4	5	6	7	8	9	10	11	12	
			Productive activity		Goods				Factors		Institutions				Total
			Value added	Gross output	Domestic	Exported	Imported	Composite	Labor	Capital	Households	Government	Investment	Rest of the World	
1	Productive activity	Value added		7,168											7,168
2		Gross output			10,509	1,915									12,425
3	Goods	Domestic						10,902							10,902
4		Exported												1,915	1,915
5		Imported						2,026							2,026
6		Composite		5,256							5,048	893	1,729		12,928
7	Factors	Labor	2,370												2,370
8		Capital	4,493												4,493
9	Institutions	Households							2,370	4,493					6,863
10		Government	304		393										697
11	Savings										1,815	-195		110	1,729
12	Rest of the world						2,026								2,026
	Total		7,168	12,425	10,902	1,915	2,026	12,928	2,370	4,493	6,863	697	1,729	2,026	

To close the supply-demand circuit, note that demand is recorded in row 6, where the various demands, intermediate consumption and final demand (households, government and investment) are registered, so that supply and demand are equal.

The following accounts describe the circular flow of income, which originates in the payment to the factors of production (value added), and is distributed in columns 7 and 8 to households. These, after deducting taxes, spend their income on consumption and savings. Final demand then arises from private incomes of households, government revenues, and savings or investment.

For the purposes of this paper, what is relevant in this section is to show how the so-called composite good is defined as the result of adding the domestic supply and imports. The next section will show that the representative national consumer selects the optimal mix between these two sources, depending on his budget constraint and the prices of domestic and imported goods. In the econometric estimation, quarterly time series were used, corresponding to the national and imported supply, with their respective price indices, as defined in our SAM of 2003.

3 The Model

Given that the Armington assumption states that domestic and imported goods are often not perfect substitutes, a CES function, that allows us to model the supply of the composite, Q, is specified as

$$Q = \varphi\big[\delta D^{-\rho} + (1-\delta)M^{-\rho}\big]^{-1/\rho}, \tag{1}$$

where D is the good produced domestically, M is the imported good, φ is a scale parameter, δ is the distribution parameter and ρ is a parameter of substitution.

The consumer minimizes his expenditure, E, subject to (1):

$$\min E = P_D D + P_M M, \tag{2}$$

$$\text{subject to } \; Q = \varphi\big[\delta D^{-\rho} + (1-\delta)M^{-\rho}\big]^{-1/\rho},$$

where P_D is the price of the domestic good and P_M is the price of the imported one. The solution of the optimization problem consists in choosing D and M so that the first order conditions of problem (2) are satisfied, which may be expressed by the relation between relative demand and relative prices, given by

$$\frac{M}{D} = \left[\left(\frac{1-\delta}{\delta}\right)\frac{P_D}{P_M}\right]^{\varepsilon}, \tag{3}$$

where $\varepsilon = \frac{1}{1+\rho} > 0$ is the Armington substitution elasticity. Linearizing the above expression, the static log - log linear model can be formulated as

$$\ln\left(\frac{M}{D}\right) = \beta + \varepsilon \ln\left(\frac{P_D}{P_M}\right), \tag{4}$$

with $\beta = \varepsilon \ln\left(\frac{1-\delta}{\delta}\right)$. From here the empirical analysis starts.

4 Econometric Methodology

The final specification of the model depends on the properties of the empirical information incorporated in the imported and domestic goods and in the relative prices, and of course, of the economic and econometric evaluations. In this case, time series are available for each variable in the static model in (4), which can be formulated in its linear form as

$$\ln Y_t = \beta + \varepsilon \ln X_t + e_t, \tag{5}$$

where $Y_t = \frac{M_t}{D_t}$, $X_t = \frac{P_{Dt}}{P_{Mt}}$, e_t is the stochastic term, which is a Gaussian white noise, $t = 1, \ldots, T$ is an index that runs over the observations, and T is the total number of them.

First, tests of stationarity are conducted and, where appropriate, the order of integration of the variables $\ln Y_t$ and $\ln X_t$ is determined.[3] When these variables are stationary, $I(0)$, the most appropriate model is the Partial Adjustment Model (PAM), which can be specified as

$$\ln Y_t = \beta + \varepsilon_1 \ln X_t + \varepsilon_2 \ln Y_{t-1} + e_t, \tag{6}$$

with the advantage that this model is dynamic and provides the short and long run Armington elasticities. In it, ε_1 is the short run elasticity, and the long run elasticity is given by

$$\varepsilon_{LP} = \frac{\varepsilon_1}{1-\varepsilon_2}. \tag{7}$$

If the variables in log-levels are not $I(0)$, their orders of integration are determined, and they are analyzed to determine if they are cointegrated only if their orders of integration are the same.[4] If cointegration of variables is accepted, an Error Correction Model (ECM) is formulated [14].[5] As in the linear model presented in (5), there are

[3] This is stationary of the second order, or covariance stationary.

[4] According to Engle and Granger [10], a set of variables is cointegrated if they have the same order of integration, $I(d)$, $d > 0$, and if there is a linear combination of them that is $I(d-b)$ (its order of integration is less than d). This linear combination is the long run relation. In this manner, the concept of cointegration refers to the existence of long run relationships between variables, so that even if these increase (or decrease), they do so in a completely synchronized way.

[5] The most used ECM formulation is

$$\Delta \ln Y_t = \beta + \varepsilon_1 \Delta X_t + \alpha\left[\ln Y_{t-1} - \gamma ln X_{t-1}\right] + e_t,$$

two variables, if they are cointegrated, there will be only one long run relationship, and if in addition both are $I(1)$, the ECM can be specified as

$$\varDelta \ln Y_t = \beta + \varepsilon_1 \varDelta \ln X_t + \varepsilon_2 \ln Y_{t-1} + \varepsilon_3 \ln X_{t-1} + e_t, \tag{8}$$

where ε_1 continues being the short run elasticity, and the long run elasticity is determined by

$$\varepsilon_{LP} = -\frac{\varepsilon_3}{\varepsilon_2}, \tag{9}$$

where ε_{LP} is the long run elasticity.[6] When the variables have the same order of integration, but are not cointegrated, a model in first differences of the log-levels is used:

$$\varDelta \ln Y_t = \beta + \varepsilon_1 \varDelta \ln X_t + e_t, \tag{10}$$

where, ε_1, as always, represents short run elasticity.

5 Estimation and Evaluation of the Armington Elasticities

The methodology presented previously notes that before proceeding to estimate a model, it is necessary to know the properties of the empirical information that is inserted in it. Therefore, first the stationarity of the time series in log-levels is analyzed; if they are not stationary, a transformation is sought (difference) that is stationary to obtain the order of integration. Next, we study whether the series with the same order of integration are cointegrated or not, to finally specify, estimate, and evaluate the appropriate model.

5.1 The Data

The estimation of the models considers quarterly information from INEGI for the period covered by the first quarter of 1993 to the fourth quarter of 2013, at constant prices of 2008. In the relative demand, M, represents total imports and D,

(Footnote 5 continued)
where $\varDelta$ is the difference operator, α is the speed of adjustment, ε_1 shows the short run effects, γ measures the long run effect of a change in the logarithm of relative prices on the logarithm of the ratio of imported to domestic goods. Doing some algebra, we can obtain the model in (8) with $\varepsilon_2 = \alpha$ y $\varepsilon_3 = -\alpha\gamma$. As already mentioned, ECM makes sense only when $\ln Y_t$ and $\ln X_t$ are $I(1)$ and cointegrated, so in this way, it is guaranteed that $\left[\ln Y_{t-1} - \delta ln X_t\right]$ is $I(0)$, and therefore the equation is balanced, as the stochastic term is assumed to be white noise, $I(0)$, and so are $\varDelta \ln Y_t$ and $\varDelta \ln X_t$.

[6] The difference of the logarithm of the variable Z_t, $\varDelta \ln Z_t = \ln Z_t - \ln Z_{t-1}$ is its rate of growth.

the domestic demand, which is calculated as the gross value of production, minus exports. Meanwhile, in the determination of relative prices, the corresponding price indices of D and M were used.

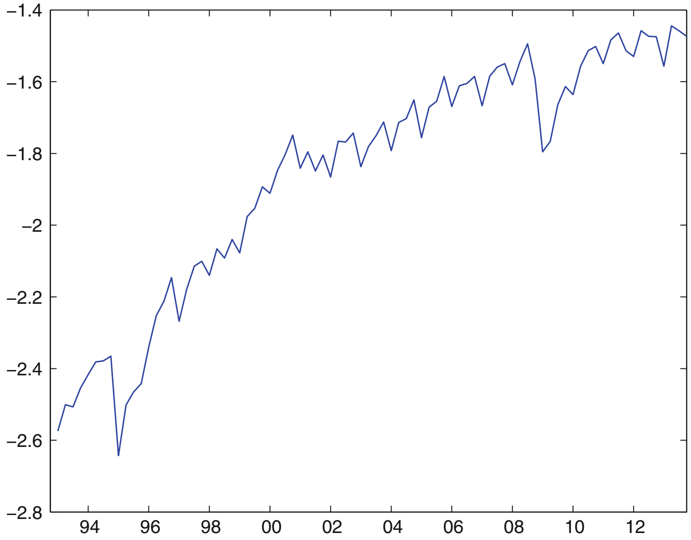


Fig. 1 Logarithm of the relative demand

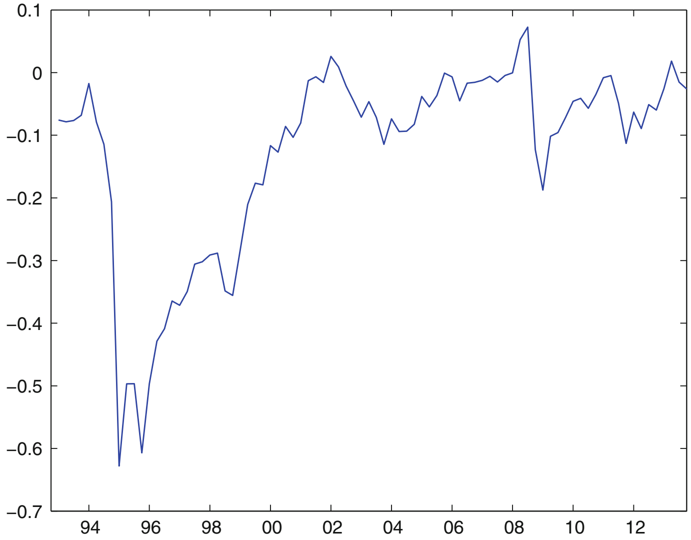


Fig. 2 Logarithm of the relative price

Table 2 Stationarity tests

Variable	ADF	PP	KPSS
$\ln\left(\frac{M_t}{YP_{Dt}}\right)$	−3.1542	−2.7090	1.0607
	(0.0569)	(0.0768)	
$\ln\left(\frac{P_{YDt}}{P_{Mt}}\right)$	−1.9087*	−2.0087*	0.6495
	(0.3269)	(0.2827)	
$\Delta\ln\left(\frac{M_t}{YP_{Dt}}\right)$	−3.3331	−13.8929	0.5000
	(0.0168)	(0.0001)	
$\Delta\ln\left(\frac{P_{YDt}}{P_{Mt}}\right)$	−7.0275	−9.1610	0.0810∗
	(0.0000)	(0.0000)	
Critical values			
5%	−2.9012	−2.8967	0.4630
10%	−2.5879	−2.5856	0.3470

The numbers in parenthesis are the p-values and the * indicates rejection of the null hypothesis at a 5% significance level

5.2 Stationarity and Cointegration Analysis

It is important to point out that regression analysis, in the presence of integrated variables, can lead to spurious relationships [12], so it is necessary to check whether the model variables are stationary; that is to say, if their mean and unconditional variance are time-invariant, and if the unconditional covariance is equal for couples of variables with the same distance in time. In case some of these properties are not satisfied, differences are applied to see the possibility of obtaining a new variable that will satisfy the properties.

Figures 1 and 2 show that neither the logarithm of the relative demand, nor of the relative price, is stationary, both show tendency. Although this is evidence of nonstationarity, statistical tests to support this fact must be performed. Here, the Augmented Dickey and Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are carried out to analyze the stationarity of the variables of the model.[7] All these tests show that the first difference of log-level of demand and relative prices are stationary at a 5% significance level and, therefore, the logarithms of these variables are $I(1)$ (Table 2).

According to the above results, the two variables of the model are $I(1)$ and, therefore, growth rates of the relative demand (substitution rate) and of the prices are $I(0)$, which means that the variables $\Delta\ln Y_t = \Delta\ln\left(\frac{M_t}{D_t}\right)$ and $\Delta\ln X_t = \Delta\ln\left(\frac{P_{Dt}}{P_{Mt}}\right)$, are stationary and, thus, there is a possibility that the variables $I(1)$ are cointegrated.

[7] In the ADF and PP tests ([6, 20], respectively), the null hypothesis is non-stationarity or equivalently, H_0: Unit root, while in the KPSS [19], H_0: No unit root (Stationarity).

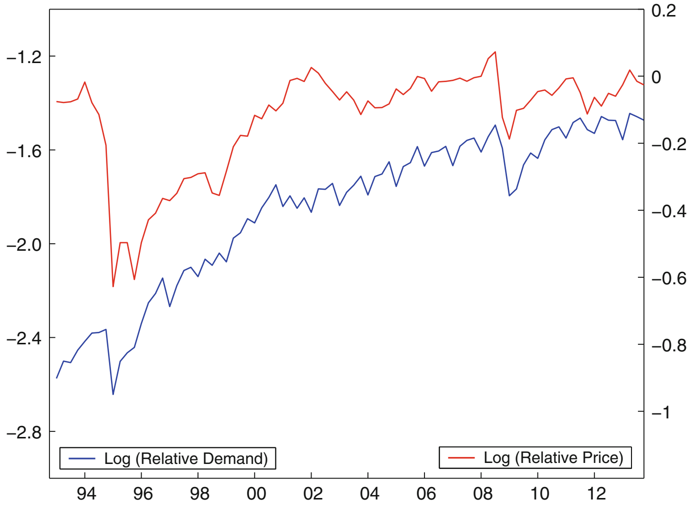


Fig. 3 Relative demand and relative price (logarithms)

Table 3 Stationarity tests of the residuals

Variable	ADF	PP	KPSS
$\hat{u}_t$	−2.8471	−3.4496	0.3154
	(0.0564)	(0.0119)	
Critical values			
5%	−2.8987	−2.8967	0.4630
10%	−2.5866	−2.5856	0.3470

The numbers in parentheses are the p-values

In Fig. 3 one observes that there could actually be a long run relationship between the logarithm of relative demand and the logarithm of the relative price, as they show very synchronized behavior. However, the graphical evidence is not sufficient to guarantee the existence of cointegration. This is confirmed or refuted by the Engle and Granger test [10] and/or the Johansen test [17], in its two versions: the maximum eigenvalue and trace. All these tests incorporate a tendency because of the dynamics of the series.

In the Engle and Granger test, one should ensure that the residuals of the regression

$$\ln Y_t = \delta_1 + \delta_2 t + \delta_3 \ln X_t + u_t, \qquad (11)$$

are stationary, for which the ADF, PP and KPSS tests are carried out. The results, shown in Table 3, do not reject the stationarity of the residuals, given by

$$\widehat{u}_t = \ln Y_t - \widehat{\ln Y_t} = \ln Y_t - \hat{\delta}_1 - \hat{\delta}_2 t - \hat{\delta}_3 \ln Y_t$$

Table 4 Johansen tests

Test of the maximum eigenvalue				
H_0	H_1	λ_{Max}	Critical value*	$p-$ value
$r=0$	$r=1$	23.7713	14.2646	0.0012
$r\leq 1$	$r=2$	8.6722	3.8415	0.0032
Trace test				
H_0	H_1	λ_{Trace}	Critical value*	$p-$ value
$r=0$	$r=1$	32.4436	15.4947	0.0001
$r\leq 1$	$r=2$	8.6722	3.8414	0.0032

r is the number of relations of cointegration
$*$ 5% significance level

For their part, the versions of the Johansen test presented in Table 4 also provide evidence that the variables are cointegrated, since in both, in the second iteration the hypothesis which establishes the existence of a cointegration relationship, is not rejected.

5.3 Estimation of the Armington Elasticities

Since the variables $\ln Y_t$ and $\ln X_t$ are $I(1)$ and are cointegrated, the appropriate model to estimate the Armington elasticities is the Error Correction Model (ECM) given by

$$\begin{aligned}\Delta \ln Y_t = \alpha_1 + \varepsilon_1 \Delta \ln X_t + \alpha_2\big(&\ln Y_{t-1} - \hat{\delta}_1 - \hat{\delta}_2(t-1)\\ &-\hat{\delta}_3 \ln X_{t-1}\big) + e_t,\end{aligned} \tag{12}$$

where the term within parenthesis is the residual of the model in (11) and based on it, the least square estimators of α_1, α_2 and ε_1 are obtained. Once terms have been associated, the model can be expressed as follows

$$\Delta \ln Y_t = \beta_1 + \beta_2 t + \varepsilon_1 \Delta \ln X_t + \varepsilon_2 \ln Y_{t-1} + \varepsilon_3 \ln X_{t-1} + e_t, \tag{13}$$

where $\beta_1 = \alpha_1 - \alpha_2(\hat{\delta}_1 - \hat{\delta}_2)$, $\beta_2 = -\alpha_2\hat{\delta}_2$, $\varepsilon_2 = \alpha_2$ and $\varepsilon_3 = -\alpha_2\hat{\delta}_3$, the last specification is the version with tendency of the ECM proposed in (8). In the previous models, Armington short run and long run elasticities are, respectively,

$$\varepsilon_1 = \delta_3 \text{ and } \varepsilon_{LP} = -\frac{\varepsilon_3}{\varepsilon_2}.$$

The ECM estimation was carried out following the Engle and Granger procedure, which consists of two stages. The first is to verify whether the model residuals in (11) are stationary. If so, they proceed to the second stage, where the model proposed in (12) is estimated using the residuals of the regression in (11). The estimation made here is done by adding dichotomous variables (dummies) to the model in (13) to account for significant changes in the level of the series in 1995 and in 2008–2009, to reflect the impact of the 1994–1995 Mexican crises and world crises, respectively, that significantly impacted the behavior of demand and relative prices (see Figs. 1 and 2).[8]

In order to obtain statistically efficient estimation and inference, one may test the weak exogeneity of $\Delta \ln X_t$ for the parameter(s) of interest.[9] Engle and Granger [10] argue that a simple way to check the weak exogeneity of $\Delta \ln X_t$ for the parameters of interest is to estimate an ECM for $\Delta \ln X_t$, and test the significance of the error correction term, using a traditional t-test. The weak exogeneity of $\Delta \ln X_t$ is not rejected for the long-run parameter, this means that the speed of adjustment coefficient appears as insignificant in the ECM for $\Delta \ln X_t$ (see Appendix A). Thus, one can conclude that $\Delta \ln X_t$ may be considered as weakly exogenous for the long run parameter.

The first stage of the Engle-Granger procedure was performed when doing the cointegration test, in which the stationarity of the residuals was verified, so we proceed to the second stage, obtaining the estimated model

$$\begin{aligned}\widehat{\Delta \ln Y_t} = & \underset{(0.0053)}{0.0270} + \underset{(0.0596)}{0.534}\Delta \ln X_t - \underset{(0.0094)}{0.095}d_{1t} + \underset{(0.0363)}{0.036}d_{2t} \\ & \underset{(0.0520)}{-0.052}\left(\ln Y_{t-1} + 2.004 - 0.0001t - 0.719 \ln X_{t-1}\right),\end{aligned}$$

where d_1 and d_2 are dummy variables and the figures in parentheses are the standard errors. The estimated parameters have the expected signs and magnitudes, are significant, and in general, the econometric evaluation is appropriate (Table 5 in the Appendix B).[10] The Armington short and long run elasticities are given by

$$\varepsilon_1 = 0.534 \text{ and } \varepsilon_{LP} = 0.719.$$

[8] Given the behavior of the log of the demand and of the relative prices, the dummies are defined as

$$d_{1t} = \begin{cases} 1, & t = 1995:1 \text{ to } 1995:4, \\ 0, & \text{in the other quarters,} \end{cases}$$

and

$$d_{2t} = \begin{cases} 1, & t = 2008:1 \text{ to } 2009:3, \\ 0, & \text{in the other quarters,} \end{cases}$$

where 1995:1 indicates the first quarter of 1995 and the other periods are defined in an analogous manner.

[9] The importance of the concept of exogeneity in a conditional econometric model has been pointed out particularly well in [9].

[10] Given that we have a cointegration relationship with $I(1)$ variables, we apply Granger causality test. The null hypothesis of no-causality from $\Delta \ln X_t$ to $\Delta \ln Y_t$ is rejected, but it is not rejected from $\Delta \ln Y_t$ to $\Delta \ln X_t$. Therefore, $\Delta \ln X_t$ is strongly exogenous, since $\Delta \ln X_t$ is weakly exogenous and $\Delta \ln Y_t$ is not Granger causing $\Delta \ln X_t$.

respectively, which implies that relative demand is inelastic in both the short and long run; that is, changes in the relative prices of domestic and imported goods do not have a substantial effect on the relative demand for imports to national goods. Furthermore, the long run elasticity obtained is larger than the short run elasticity, largely due to the longer adjustment time.

6 Conclusions

In this paper, short and long run Armington elasticities have been estimated for the Mexican economy. The specification of the final model depended on a detailed assessment of the assumptions of the econometric model. The estimations were made using an Error Correction Model, since the two variables of the model, logarithm of relative demand and logarithm of relative prices, are integrated of order one and are cointegrated. This model has the advantage of providing both the short and long run elasticity estimates, and the estimation is adequate, since on the one hand, elasticities have the expected signs and magnitudes, and on the other, the estimated parameters are individually and jointly significant, and the residuals satisfy the assumptions underlying the stochastic terms of the theoretical econometric model.

In the estimation of the Error Correction Model, the long run Armington elasticity is greater than the short run due to the longer adjustment time. Thus, both elasticities suggest that domestic and imported goods are poor substitutes in Mexico, as in IRBC models. In future research, Armington elasticities will be estimated with disaggregated data, and the discrepancies between the elasticities obtained with aggregated and disaggregated data will be observed. Thus, trade flows in Mexico could be better understood, either using IRBC models, or CGE models.

Appendix A. Weak Exogeneity Test

The estimated ECM for $\varDelta \ln X_t$ is

$$\varDelta \ln X_t = \underset{(0.0077)}{0.0050} - \underset{(0.1077)}{0.148}\varDelta \ln Y_t + \underset{(0.0772)}{0.087}\big(\ln X_{t-1} - 0.467 - 0.321 \ln Y_{t-1}\big)$$

where the figures in parentheses are the standard errors. The long-run coefficient is insignificant, therefore weak exogeneity is not rejected and one can conclude that $\varDelta \ln X_t$ may be considered as weakly exogenous for the long run parameter.

Table 5 Diagnostic tests for the ECM

Test	H_0	Statistics	$p-$ value
Normality			
-Jarque-Bera	Normality	5.4462	0.0656
Autocorrelation	No autocorrelation		
-Lungj-Box			
1 lag		0.0291	0.8650
4 lag		1.3635	0.8510
12 lag		10.559	0.5670
32 lag		31.846	0.4740
-Breusch-Godfrey	No autocorrelation		
1 lag		0.2860	0.8661
4 lag		0.3164	0.8661
Heteroskedasticity	Homoskedasticity		
-White		0.7369	0.5696
-White terms Crossed		1.0002	0.4551
-Breusch-Pagan-Godfrey		1.4618	0.2219
Correct Specification	Linearity		
-RESET		1.3488	0.1814

Appendix B. Diagnostic Tests

Table 5 presents the results of diagnostic tests of the estimated model for Armington elasticities. According to the results, it can be concluded that their residuals are a good proxy for the stochastic term of the theoretical econometric model because they are normal, not autocorrelated and homoskedastic, and also, the linear model specification is correct (Table 5).

References

1. Anderson, J.E., van Wincoop, E.: Trade costs. J. Econ. Lit. **42**, 691–751 (2004)
2. Armington, P.S.: Theory of demand for products distinguished by place of production. Int. Monet. Fund Staff Pap. **16**, 159–176 (1969)
3. Backus, D.K., Kehoe, P.J., Kydland, F.E.: Dynamics of the trade balance and the terms of trade: the j-curve? Am. Econ. Rev. **84**, 84–103 (1994)
4. Bergin, P.R.: How well can the new open economy macroeconomics explain the exchange rate and current account? J. Int. Money Finance **25**, 675–701 (2006)
5. Crucini, M.J., Davis, J.S.: Distribution capital and the short and long-run import demand elasticity, NBER Working Paper No. 18753 (2013)
6. Dickey, D., Fuller, W.: Distribution of the estimators for autoregressive time series with unit root. J. Am. Stat. Assoc **74**, 427–431 (1979)

7. Engel, C., Wang, J.: International Trade in Durable Goods: Understanding Volatility, Cyclicality, and Elasticities, NBER Working Paper No. 13814 (2008)
8. Engel, C., Wang, J.: International trade in durable goods: understanding volatility, cyclicality, and elasticities. J. Int. Econ. **83**, 37–52 (2011)
9. Engle, R.F., Hendry, D.F., Richard, J.F.: Exogeneity. Econometrica **51**, 277–304 (1983)
10. Engle, R.F., Granger, C.W.J.: Cointegration and error correction: representation, estimation and testing. Econometrica **55**, 251–276 (1987)
11. Gallaway, M.P., McDaniel, C.A., Rivera, S.A.: Short-run and long-run industry-level estimates of U.S. armington elasticities. N. Am. J. Econ. Financ. **14**, 49–68 (2003)
12. Granger, C.W.J., Newbold, P.: Spurious regressions in econometrics. J. Econom. **2**(2), 111–120 (1974)
13. Heathcote, J., Perri, F.: Financial autarky and international business cycles. J. Monet. Econ. **49**(3), 601–628 (2002)
14. Hendry, D.F., Pagan, A.R., Sargan, D.J.: Dynamic specification. In: Griliches, Z., Intrilligator, M.D. (eds.) Handbook of Econometrics, vol. 3. North-Holland, Amsterdam (1984)
15. Hertel, T., Hummels, D., Ivanic, M., Keeney, R.: How confident can we be of cge-based assessments of free trade agreements? Econ. Model. **24**, 611–635 (2007)
16. Hooper, P., Johnson, K., Marquez, J.: Trade Elasticities for the G-7 Countries. Princeton Studies in International Economics No. 87 (2000)
17. Johansen, S.: Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica **59**, 1551–1580 (1991)
18. Kose, M.A., Yi, K.M.: Can the standard international business cycle model explain the relation between trade and co-movement? J. Int. Econ. **68**, 267–295 (2006)
19. Kwiatkowski, D., Philips, P., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationary against the alternative of unit root. J. Econom. **54**, 159–178 (1992)
20. Phillips, P., Perron, P.: Testing for a unit root in time series regression. Biometrika **75**(2), 335–346 (1988)
21. Ruhl, K.J.: The International Elasticity Puzzle, University of Texas at Austin (2008)
22. Sobarzo, H.: The gains for Mexico from a North American free trade agreement - an applied general equilibrium assessment. In: Francois, J.F., Shiells, C.R. (eds.) Modeling Trade Policy: Applied General Equilibrium Assessments of NAFTA. Cambridge University Press (1994)

Mathematical Modelling for Wave Drag Optimization and Design of High-Speed Aircrafts

Can Citak, Serkan Ozgen and Gerhard Wilhelm Weber

Abstract Supersonic flight has been the subject of last half century. Both civil and defence projects have been running to design an aircraft to fly faster than speed of sound. Developing technology and increasing experience of design leads to faster, fuel efficient, hence, ecological, long-ranged aircrafts. These vehicles make people live easy by shortening travel time, perform missions with powerful defence aircrafts and helping explore space. Aerodynamic design is the main argument of the high speed aircrafts improvement. Having less supersonic drag force, which is greater than the double of subsonic case for conventional aircraft, is the ultimate goal of the aircraft designers at supersonic speed. In this chapter, an aerodynamic characteristics of the entire configuration is optimized in order to reach this aim. Moreover, solver algorithm is validated with computational fluid dynamics simulations for different geometries at various speeds. The objective of this study is to develop a program which optimizes wave drag coefficient of high speed aircrafts by numerical methods.

Keywords Supersonic flight · Wave drag · Optimization · Area rule

1 Wave Drag Definition

Designing an aircraft with the ability of flying faster than the speed of sound was the purpose of most aircraft designers in the past decades in order to reduce travel time and research space. Both aims require ultimate design configurations for definite missions. Unlike the subsonic design, the supersonic region has struggles to deal with in order to reach this aim. The major part of this problem is about the huge drag

C. Citak (✉) · S. Ozgen · G.W. Weber
METU, Ankara, Turkey
e-mail: ccitak@ae.metu.edu.tr

S. Ozgen
e-mail: sozgen@ae.metu.edu.tr

G.W. Weber
e-mail: gweber@metu.edu.tr

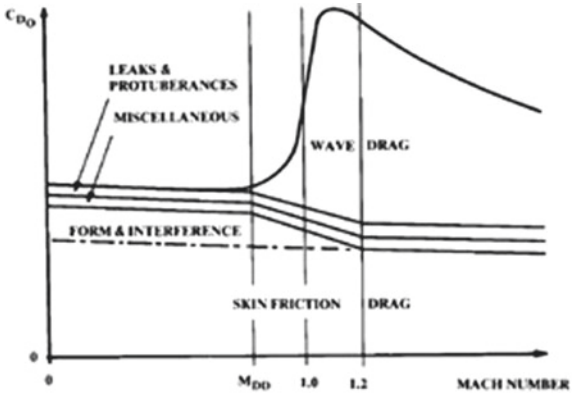
A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_7

force occurring when compared to subsonic speed. Thus, aircraft designers aware of these drawbacks were in a need of making modifications to their design. For example, re-entry of an spacecraft which is directly related to drag force must be considered as one of the critical issue of the overall design process.

Wave drag can be described as the major part of the force resisting aircraft motion at supersonic speed. It depends on the velocity of the aircraft, wing area, air density and drag coefficient which are related to complete configuration of the aircraft. The main purpose in an aircraft design is generally to reduce drag to minimum level. On the other hand, drag force is beneficial for some extreme cases, such as the utilization of parachute for short distance landing. Drag is mainly classified as drag due to lift and zero lift drag. The work represented in this chapter mainly concentrates on the wave drag (zero lift). Temperature, pressure, aircraft velocity and the shape of the configuration affects the magnitude of the wave drag. When supersonic free stream reaches an obstacle, shock wave occurs which increases the density and pressure of the flow. In other words, the free stream Mach number, which must be greater than 1 for shock wave to occur, decreases below Mach 1 after the normal shock formation [5]. The shock wave leads to increase in entropy and reduction in total pressure. If the shock wave is inevitable, the efficiency of the shock formation can be increased in order to reduce the total increase of entropy. A wing with sweep angle and fuselage shaping can be used for this purpose. This study aims at minimizing wave drag coefficient without changing the aerodynamic characteristics of the lifting surfaces. Thus, the area distribution and the volume of the fuselage is modified to reach the minimum value of the objective function.

As seen in Fig. 1 [14], the supersonic drag of an aircraft rises 3–4 times of the subsonic case so that the drag optimization of the supersonic aircraft is the main criterion of the aerodynamic design process. The aircraft shape might be optimized despite the fact that the composition of it seems suitable for the residential of sub-

Fig. 1 Drag variation with mach number [14]



components. Nevertheless, the optimal shape of aircraft are not being implemented to the base design due to the manufacturing and sub-component constraints which give rise to additional drag. Small changes in supersonic drag could be critical. To illustrate this, on the *Concorde*, [16] it can be stated that one count drag increase ($\triangle C_d = 0.0001$) requires two passengers, out of the 90–100 passenger capacity, be taken off the North Atlantic run [16]. Additional drag components at supersonic speed are wave drag due to lift and wave drag due to volume. Wave drag due to lift vanishes as Mach number goes to one or aspect ratio goes to zero. Consequently, wave drag due to volume is investigated in this research. The behavior of the volume wave drag at various Mach numbers and different geometries are observed.

2 Far-Field Theory

Total momentum change in streamwise direction of control volume is equal to the drag of the aircraft. Inlet region is the only undisturbed flow passing through the aircraft geometry which becomes two dimensional because of the pressure effects. Thus, the momentum change between inlet and outlet regions (streamwise momentum change) is the sum of all the drag contributors. In addition, subsonic flow becomes parallel at outlet if the control volume is large enough. On the other hand, mass flows in and out from the side of the cylinder at supersonic speed due to shock and expansion wave formations [15, 17] (Fig. 2).

Total change in momentum as a result of mass flow in and out is defined as wave drag. Moreover, since the shock formation varies with the angle of attack, wave drag can change with the angle of attack as well. Therefore, wave drag is formed with wave drag due to volume and wave drag due to lift which produces the effects of wave drag variation due to lift. The drag equation is given in Eq. (1) as

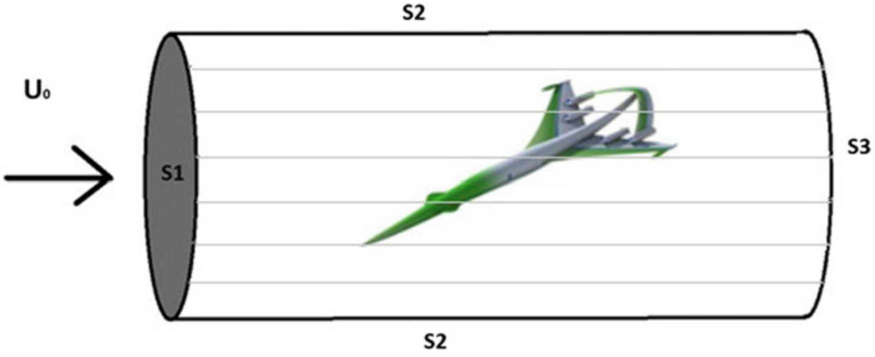


Fig. 2 Control volume representation [19]

$$\iint\limits_{S_3=S_1} (p - p_\infty)\, dS_3 - \rho_\infty U_\infty^2 \iint\limits_{S_3=S_1} \phi_x\, (1 + \phi_x)\, dS_3 - \rho_\infty U_\infty^2 \iint\limits_{S_2} \phi_x \phi_r\, dS_2 + \sum D_{misc}\ . \quad (1)$$

Miscellaneous drag consists of excrescence and base drags. If the control volume is located far enough, flow becomes two dimensional, the stream wise perturbation velocity is zero. Thus the second integral in the general drag formula becomes zero as

$$\rho_\infty U_\infty^2 \iint\limits_{S_3=S_1} \phi_x\, (1 + \phi_x)\, dS_3 = 0\ , \quad (2)$$

and the gauge pressure is formulated as

$$p - p_\infty\ = -\frac{1}{2}\rho_\infty U_\infty^2\, (\phi_x^2 + \phi_z^2)\ . \quad (3)$$

Since the viscosity effects are neglected, the total inviscid drag equation can be written as shown below:

$$D = -\rho_\infty U_\infty^2 \iint\limits_{S_2} \phi_x \phi_r\, dS_2 + \frac{1}{2}\rho_\infty U_\infty^2 \iint\limits_{S_2} (\phi_y^2 + \phi_z^2)\, dS_3\ . \quad (4)$$

Wave drag can be calculated directly from mass flow change at side surface of the control volume. Perturbation velocities in the first integral give the velocity change in side direction. As these are multiplied with the density and the square of free stream velocity, the total wave drag could be obtained. The wave drag formula is given in Eq. (5):

$$D_w = -\rho_\infty U_\infty^2 \iint\limits_{S_2} \phi_x \phi_r\, dS_2\ . \quad (5)$$

Farfield linear theory has positive and negative characteristics. First, it is simply used for calculations. In addition, singularities can be overcome without sophisticated numerical methods; i.e., pressure calculations at leading edge. As shown in Eq. (4), induced drag can be separated from wave drag by using far field linear theory which provides pure wave drag calculation. Thus, it is useful for area rule optimization with respect to wave drag. Since the volume of the aircraft is the only contributor to drag formula, aircraft geometry can be directly related to the wave drag. Hence, aircraft area distribution can be modified in order to minimize wave drag. On the contrary, the theory does not reflect physics of the flow completely. Therefore, aircraft design could be validated with other methods to ensure behavior of the flow over aircraft surface [2].

2.1 Formula Transformation

Conventional form of the wave drag is given as

$$D_w = -\frac{1}{2\pi} \int_0^1 \int_0^1 S''(x)S''(y)\ log|x-y|\ dxdy\ . \tag{6}$$

Two problems arise in the calculation of the formula given above. Firstly, singularity occurs where the longitudinal locations of the aircraft become identical. Secondly, numerical precision strongly depends on the differentiation method used and the degree of accuracy. Thus, a sensitivity analysis is effective for calculation of wave drag force. Two conditions must be satisfied for the method used to obtain wave drag:

1. The first derivative of the area distribution is continuous along longitudinal direction of aircraft.
2. The first derivatives of the area distribution at nose and rear regions are equal to zero:

$$S'(0) = S'(L) = 0\ , \tag{7}$$

where L represents the length of aircraft. When the conditions explained above are satisfied, the first derivative of the area distribution can be transformed to the Fourier sine series as,

$$x = \frac{1}{2}\ (1 - \cos\theta)\ , \tag{8}$$

where θ varies between 0 and π:

$$\theta = \cos^{-1}(1 - 2x)\ . \tag{9}$$

Subsequently, we refer to Eqs. (8) and (9) implicitly. Then the first derivative distribution is given by

$$S'(x) = \sum_{r=1}^{\infty} a_r\ \sin r\theta\ , \qquad 0 \le\ x \le 1\ , \tag{10}$$

where the coefficient is written as

$$a_r\ = \sum_{r=1}^{\infty} \frac{2}{\pi} \int_0^{\pi} S'(x) r\theta d\theta\ . \tag{11}$$

The area distribution of the aircraft is obtained by integrating the Eq. (10) as

$$S(x) = \sum_{r=1}^{\infty} a_r \int_0^{\pi} \sin r\theta dx \ . \tag{12}$$

Equation (9) is integrated and substituted into Eq. (11) by using the derivative of Eq. (8):

$$dx = -\frac{1}{2} \ \sin\theta \ d\theta \ , \tag{13}$$

hence,

$$S(x) = \frac{1}{2} \sum_{r=1}^{\infty} a_r \int_0^{\pi} \sin r\theta \sin\theta d\theta,$$

$$= a + \frac{1}{4} a_1 \ (\theta - \frac{1}{2} \ \sin 2\theta) + \frac{1}{4} \sum_{r=2}^{\infty} a_r [\frac{\sin(r-1)\theta}{r-1} \ - \frac{\sin(r+1)\theta}{r+1} \ ],$$

$$= a + \frac{1}{4} a_1 \ \theta + \frac{1}{4} \sum_{r=1}^{\infty} (a_r - a_r - 1) \sin r\theta \ . \tag{14}$$

By using Eq. (13), the second derivative of the area distribution is obtained and inserted into Eq. (6) as

$$D_w = \frac{1}{2} \int_0^{\pi} \sum_{r=1}^{\infty} a_s \ \sin s\theta d\theta,$$

$$= \frac{1}{2} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} r a_r a_s \int_0^{\pi} \sin r\theta \sin s\theta d\theta,$$

$$= \frac{\pi}{4} \sum_{s=1}^{\infty} r a_r^2 \ . \tag{15}$$

Gradient-based optimization method is used in order to obtain the area distribution which has minimum wave drag force. Since the accuracy of the gradient calculation strictly depends on the smoothness of the objective function and constraints.

3 Mathematical Modelling

3.1 Cross-Sectional Area Calculation

Greens theorem is used for the calculation of cross-sectional area [10]. The incremental area dA is defined as

$$dA = dxdy \ . \tag{16}$$

It states that area A of a closed region D can be represented as

$$A = \iint\limits_{D} dA \ . \tag{17}$$

Furthermore, M and L are functions having continuous partial derivatives defined by the boundaries of D:

$$\frac{\partial M}{\partial x} - \frac{\partial L}{\partial y} = 1 \ . \tag{18}$$

The area of A is given as

$$A = \oint\limits_{C} (Ldx + Mdy) \ . \tag{19}$$

The final form of the area formula can be written as

$$A = \frac{1}{2} \oint\limits_{C} (-ydx + xdy) \ . \tag{20}$$

Area computation for each cross-section is necessary as being inputs to the solver, since the shape of the cross-sections are arbitrary with variable number of points. Equation (19) is used to calculate this area. Figure 3 indicates the arbitrary shaped cross section:

$$S = \frac{1}{2} \sum_{i=1}^{n-1} (y_i x_{i+1} - y_{i+1} x_i) \ . \tag{21}$$

3.2 Fourier Transformation

The Fourier transformation methodology is defined as fitting the data set or any type of the polynomial to sinusoidal function(s). General formulation for the polynomial curve fitting is written as

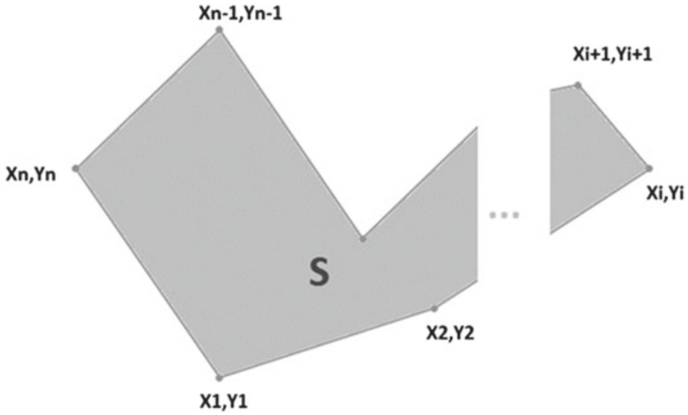


Fig. 3 Arbitrary shaped area

$$y = a_0 + a_1 x + a_1 x^2 + \cdots + a_m x^n \ . \tag{22}$$

The residual is calculated as

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x - a_1 x^2 - \cdots - a_m x^n)^2 \ ; \tag{23}$$

this fit of the curve accuracy has to be optimized. Thus, gradients of the residual is zero when the curve fitting represents the data set successfully. The gradients are given by

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x - a_1 x^2 - \cdots - a_m x^n) \ , \tag{24}$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x - a_1 x^2 - \cdots - a_m x^n) \ , \tag{25}$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^{n} x_i^2 (y_i - a_0 - a_1 x - a_1 x^2 - \cdots - a_m x^n) \ , \tag{26}$$

$$\vdots$$

$$\frac{\partial S_r}{\partial a_n} = -2 \sum_{i=1}^{n} x_i^n (y_i - a_0 - a_1 x - a_1 x^2 - \cdots - a_m x^n) \ . \tag{27}$$

The coefficients are obtained by equating and solving the gradient equations as

$$(n)a_0 + (\sum x_i)a_1 + (\sum x_i^2)a_2 + \cdots + (\sum x_n^m)a_m = \sum y_i \ , \tag{28}$$

$$(\sum x_i)a_0 + (\sum x_i^2)a_1 + (\sum x_i^3)a_2 + \cdots + (\sum x_n^{m+1})a_m = \sum x_i y_i \ , \tag{29}$$

$$(\sum x_i^2)a_0 + (\sum x_i^3)a_1 + (\sum x_i^4)a_2 + \cdots + (\sum x_n^{m+2})a_m = \sum x_i^2 y_i \ , \tag{30}$$

$$\vdots$$

$$(\sum x_i^n)a_0 + (\sum x_i^{n+1})a_1 + (\sum x_i^{n+2})a_2 + \cdots + (\sum x_n^{m+n})a_m = \sum x_i^m y_i \ . \tag{31}$$

The same approach can be used for the Fourier transformation. The polynomial function can be changed into the sinusoidal variables in order to fit the Fourier transformation to data set. Equation (30) represents first order Fourier model. Application of the transformation is presented as follows:

$$y = a_0 + a_1 \cos(\omega t) + b_1 \sin(\omega t) \ . \tag{32}$$

The residual of the model is given as

$$S_r = \sum_{i=1}^{n} (y_i - a_0 + a_1 \cos(\omega t) + b_1 \sin(\omega t))^2 \ , \tag{33}$$

the gradients of the residual S_r are represented by

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - A_0 + A_1 \cos(\omega t) + B_1 \sin(\omega t)) \ , \tag{34}$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^{n} \cos(\omega t)(y_i - A_0 + A_1 \cos(\omega t) + B_1 \sin(\omega t)) \ , \tag{35}$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^{n} \sin(\omega t)(y_i - A_0 + A_1 \cos(\omega t) + B_1 \sin(\omega t)) \ . \tag{36}$$

A necessary condition for success of convex curve fitting operation is that the gradient equations are equal to zero. Then, the unknown coefficients are obtained from solutions of set of equations. In this section, first derivative of the cross-sectional area distribution is transformed into Fourier sine function. The reason of this process is that the first derivative of cross-sectional area distribution must be continuous according to the wave drag calculation methodology. Equation (35) represents the open form of the sine function. In addition, smoothness is one of the most important criteria for minimization procedure. Thus, representation of real cross-sectional area distribution must be accurate enough [2, 4, 12]. Furthermore, value of error function shown in Eq. (31), does not reduce linearly. Therefore, to keep CPU at a certain level and to obtain valid representation, fourth-order sine functions are chosen;

$$\begin{aligned} y = a_0 + a_1 \cos(x) + b_1 \sin(x) + a_2 \cos(2x) + b_2 \sin(2x) \\ + a_3 \cos(3x) + b_3 \sin(3x) + a_4 \cos(4x) + b_4 \sin(4x) \ . \end{aligned} \tag{37}$$

The function S'_r gives the difference between discrete response data and the approximated function.

$$S'_r = \sum_{i=1}^{N} (y_i - y)^2 \ , \tag{38}$$

we require:

$$\frac{\partial S'_r}{\partial a_0}, \frac{\partial S'_r}{\partial a_1}, \frac{\partial S'_r}{\partial a_2}, \frac{\partial S'_r}{\partial a_3}, \frac{\partial S'_r}{\partial a_4}, \frac{\partial S'_r}{\partial b_1}, \frac{\partial S'_r}{\partial b_2}, \frac{\partial S'_r}{\partial b_3}, \frac{\partial S'_r}{\partial b_4} = 0 \ . \tag{39}$$

3.3 Point Update

Updating the points after optimization step is the final operation of the program. Simple methodology is used for this work. Initial cross-sectional area magnitude at ith location S_{init_i} is calculated as explained in the previous section. Then, optimal cross-sectional area magnitude S_{opt_i} is obtained after the optimization process. The ratio R_i is defined by

$$S_{opt_i} = S_{init_i} + \Delta S_i \ , \tag{40}$$

$$R_i = \sqrt{\frac{S_{opt_i}}{S_{init_i}}} \ . \tag{41}$$

With respect to initial X and Y locations, jth order of ith section; $P_{X_{ijinit}}$ and $P_{Y_{ijinit}}$ are updated as follows:

$$Px_{ij_{opt}} = R_i \cdot Px_{ij_{init}} \ , \tag{42}$$

$$Py_{ij_{opt}} = R_i \cdot Py_{ij_{init}} \ . \tag{43}$$

All cross-sections except for the control surfaces are updated as explained above. The idea behind the use of ratio R_i is that the slope of the points belonging to the same cross-section is kept constant. The slopes of $P_{ij_{init}}$ and $P_{ij_{opt}}$ can be written as,

$$C_{ij_{init}} = \frac{Py_{ij_{init}}}{Px_{ij_{init}}}, \quad C_{ij_{opt}} = \frac{Py_{ij_{opt}}}{Px_{ij_{opt}}} \tag{44}$$

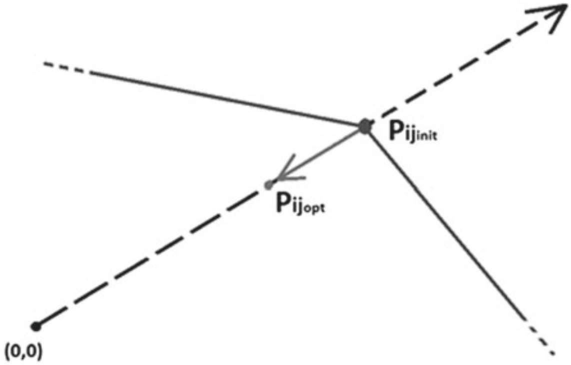


Fig. 4 Point update representation

and

$$C_{ij_{opt}} = C_{ij_{init}} = \frac{Py_{ij_{init}} \cdot R_i}{Px_{ij_{init}} \cdot R_i} \ . \tag{45}$$

To illustrate them, Fig. 4 represents the methodology behind the point update. Assuming that the final cross-sectional area is less than the initial area, then, the slope of the point can be kept constant, and updated with respect to ratio of optimal and initial cross-sectional area. Thus, the shape of the geometry is protected, which means that the initial conceptual design criteria is protected.

Furthermore, some additional steps must be investigated for non-symmetric cases. The center of the cross section must be found. Theoretically, x_c and y_c are the central locations of jth cross-section:

$$x_{c_j} = \frac{1}{n} \sum_{i=1}^{n} x_{init} \ , \tag{46}$$

$$y_{c_j} = \frac{1}{n} \sum_{i=1}^{n} y_{init} \ , \tag{47}$$

the slopes of each point in non-symmetric cross-section are

$$C_{ij_{init}} = \frac{y_{init} - y_{c_j}}{x_{init} - x_{c_j}} \ . \tag{48}$$

Since the slope is kept constant, two unknowns and two equations arise:

$$C_{ij_{opt}} = \frac{y_{opt} - y_{c_j}}{x_{opt} - x_{c_j}} \ , \tag{49}$$

$$\sqrt{(x_{init} - x_{cj})^2 + (y_{init} - y_{cj})^2} \cdot R = \sqrt{(x_{opt} - x_{cj})^2 + (x_{opt} - x_{cj})^2} \ . \tag{50}$$

Coordinates of the optimal form of the cross-sectional area distribution can be obtained by solving Eqs. (47) and (48). This approach provides an analysis of more realistic configurations.

4 Theory of Lagrange Multipliers

The methodology of Lagrange multiplier is employed for the constrained optimization. This part presents the method used for minimization of wave drag coefficient. A general formulation can be represented as

$$\min f(x) \quad \text{subject to} \begin{cases} c_i(x) = 0, & i \in \varepsilon, \\ c_i(x) \geq 0, & i \in I, \end{cases}$$

where the both objective function and constraints are smooth, real-valued functions. $i \in \varepsilon$ are the equality constraints, $i \in I$ are the inequality constraints. There are more than one local solutions for an objective function both for constrained and unconstrained cases. Smoothness of the objective functions and constraints is critical for the global convergence. Furthermore, sharp changes of these functions might mislead the search direction. To avoid that, the functions having sharp edges could characterized which can be represented with collection of smooth functions. For a simple example, Lagrangian function for one equality constraint is shown as

$$L(x, \lambda) = f(x) - \lambda_1 c_1(x) \ , \tag{51}$$

where$f(x)$ is the objective and $c_1(x)$ is the equality constraint function. The optimality condition is given as

$$\nabla_x L(x^*, \lambda_1^*) = 0, \ \text{and} \ \lambda_1^* \geq 0 \ . \tag{52}$$

Despite the fact that equation shown above is necessary for optimal solution, it is not sufficient already. It is also required that the following complementarity condition holds:

$$\lambda_1^* c_1(x^*) = 0. \tag{53}$$

Let us emphasize that, in our project, the objective function will be strictly convex, guaranteeing that our candidate solution will be a real solution. Generally, the Lagrangian function for the constrained optimization problem is defined as,

$$L(x, \lambda) = f(x) - \sum_{i=1}^{n} \lambda_i c_i(x) \ . \tag{54}$$

The active set $i \in A(x) \subseteq I$ at any feasible x is the union of the set with the indices of the active inequality constraints (where $c_i(x) = 0$ is fulfilled) [7]. Next, the linear independence constraint qualification (*LICQ*) holds since the set of active constraint gradients is linearly independent. Finally, the open form of the first-order necessary conditions is written as

$$\nabla_x L(x^*, \lambda^*) = 0 \,, \tag{55}$$

$$c_i(x^*) = 0 \,, \text{ for all } i \in \varepsilon \,, \tag{56}$$

$$c_i(x^*) \geq 0 \,, \text{ for all } i \in I \,, \tag{57}$$

$$\lambda_i^* \geq 0 \,, \text{ for all } i \in I \,, \tag{58}$$

$$\lambda_i^* c_i(x^*) = 0 \,, \text{ for all } i \in \varepsilon \cup I \,. \tag{59}$$

The multi-constrained (equality) optimization method is subsequently employed for this study. Theory of Lagrange multiplier for related subjects is explained in detail. Considering the case of objective function$f(x, y, z)$ to be minimized with respect to constraints $c_1(x, y, z)$ and $c_2(x, y, z)$. The Lagrangian function is written as

$$L(x, y, z, \lambda_1, \lambda_2) = f(x, y, z) - \lambda_1 c_1(x, y, z) - \lambda_2 c_2(x, y, z) \,, \tag{60}$$

the optimality condition is reached when,

$$\nabla f(x^*, y^*, z^*) = \lambda_1 \nabla c_1(x^*, y^*, z^*) + \lambda_2 \nabla c_2(x^*, y^*, z^*) \,. \tag{61}$$

Open form of the equations are represented as,

$$0 = L_x(x^*, y^*, z^*, \lambda_1, \lambda_2) = f_x(x^*, y^*, z^*) - \lambda_1 c_{1_x}(x^*, y^*, z^*) - \lambda_2 c_{2_x}(x^*, y^*, z^*) \,, \tag{62}$$

$$0 = L_y(x^*, y^*, z^*, \lambda_1, \lambda_2) = f_y(x^*, y^*, z^*) - \lambda_1 c_{1_y}(x^*, y^*, z^*) - \lambda_2 c_{2_y}(x^*, y^*, z^*) \,, \tag{63}$$

$$0 = L_z(x^*, y^*, z^*, \lambda_1, \lambda_2) = f_z(x^*, y^*, z^*) - \lambda_1 c_{1_z}(x^*, y^*, z^*) - \lambda_2 c_{2_z}(x^*, y^*, z^*) \,, \tag{64}$$

$$0 = L_{\lambda_1}(x^*, y^*, z^*, \lambda_1, \lambda_2) = c_1(x^*, y^*, z^*) \,, \tag{65}$$

$$0 = L_{\lambda_2}(x^*, y^*, z^*, \lambda_1, \lambda_2) = c_2(x^*, y^*, z^*) \,, \tag{66}$$

where λ_1 and λ_2 are Lagrange multipliers, "$*$" denotes the optimal condition.

4.1 Optimization Procedure

Since the area distribution is defined two different sine function which are independent, the wave drag formula is transformed into Eq. (65) (a_0 represents the *nose area* which is equal to zero):

$$D = \sum_{n=1}^{\infty} (na_n^2 + nb_n^2) \ . \tag{67}$$

The coefficients a_n and b_n above are the parameters in the Fourier transformation in Eq. (35). The permanent constraint function [9] which defines the total volume of the aircraft is defined as

$$V = \frac{1}{2} \sum_{i=1}^{k-1} (y_{i+1} + y_i) \cdot (x_{i+1} + x_i) \ , \tag{68}$$

where y represents the Fourier transformation of area distribution. The volume function is created by using a simple trapezoid rule [15]. The second constraint function is generated for keeping ith cross sectional area constant [19]. Equation (67) shows the constraint function of area

$$S_i = S_c \ . \tag{69}$$

In open form of Eqs. (66) and (67) are written as

$$\begin{aligned} C_1 = \frac{1}{8} \sum_{i=1}^{n-1} (&a_1\theta + a_2 \sin\theta + 4(a_3 - a_1)\sin 2\theta + (a_4 - a_2)\sin 3\theta + b_1\theta + b_2\cos\theta \\ &+4(b_3 - b_1)\cos 2\theta + (b_4 - b_2)\cos 3\theta) \cdot (x_{i+1} - x_i) - V = 0, \end{aligned} \tag{70}$$

$$\begin{aligned} C_2 = \big(&a_1\theta + a_2\sin\theta + 4(a_3 - a_1)\sin 2\theta + (a_4 - a_2)\sin 3\theta + b_1\theta + b_2\cos\theta \\ &+ 4(b_3 - b_1)\cos 2\theta + (b_4 - b_2)\cos 3\theta\big) - S_c = 0. \end{aligned} \tag{71}$$

Lagrangian conditions are given by

$$\begin{aligned} 0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{a_1} &= f_{a_1}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \\ \lambda_1 C1_{a_1}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) &- \lambda_2 C2_{a_1}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \end{aligned} \tag{72}$$

$$\begin{aligned} 0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{a_2} &= f_{a_2}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \\ \lambda_1 C1_{a_2}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) &- \lambda_2 C2_{a_2}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \end{aligned} \tag{73}$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{a_3} = f_{a3}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_1 C_{1_{a_3}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{a_3}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (74)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{a_4} = f_{a4}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_1 C_{1_{a_4}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{a_4}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (75)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{b_1} = f_{b1}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_1 C_{1_{b_1}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{b_1}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (76)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{b_2} = f_{b_2}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) -\lambda_1 C_{1_{b_2}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{b_2}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (77)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{b_3} = f_{b_3}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) -\lambda_1 C_{1_{b_3}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{b_3}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (78)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{b_4} = f_{b_4}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) -\lambda_1 C_{1_{b_4}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) - \lambda_2 C_{2_{b_4}}(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4), \quad (79)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{\lambda_1} = C_1 \ , \quad (80)$$

$$0 = L(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2)_{\lambda_2} = C_2 \ . \quad (81)$$

In order to reach the optimality conditions, a search direction is utilized to update iterative algorithm. The search direction is written as

$$\nabla f(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4, \lambda_1, \lambda_2) = \lambda_1 C_1(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) +\lambda_2 C_2(a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4) \ . \quad (82)$$

A convergence criterion is satisfied as soon as $\|\nabla L\|_2 \leq \varepsilon$ at the regarded point, for some given $\varepsilon > 0$.

5 Validation of the Solver with F-16 Aircraft Geometry

The solver used for the calculation of the wave drag force and coefficient is validated with CFD results. The wave drag coefficient of the F-16 aircraft at Mach 2 is obtained for comparison with Rallabhandi's result [13]. Figure 5 represents the mesh of the geometry.

Fig. 5 F-16 mesh

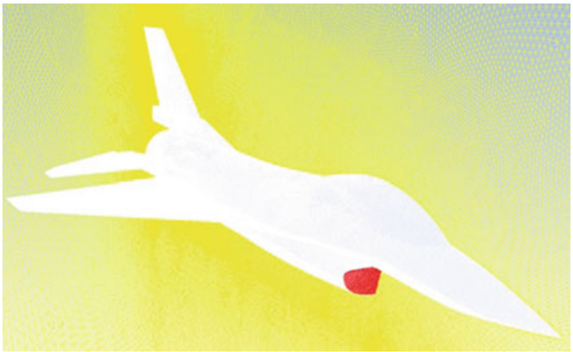


Table 1 Comparison of the results

Mach number	Rallabhandi's result C_{D_w}	Present study C_{D_w}
2.00	0.0357	0.0330

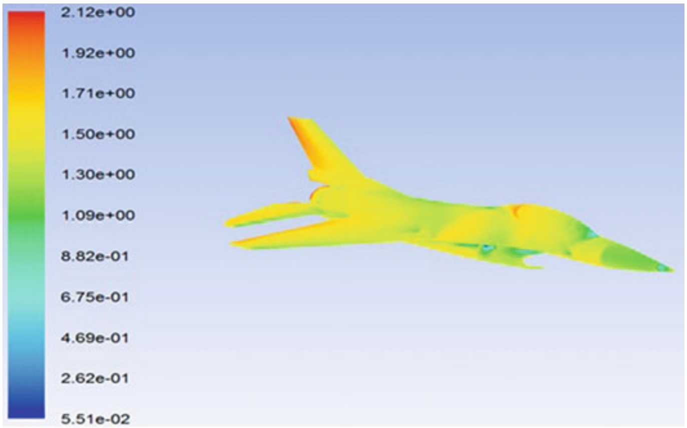


Fig. 6 Mach contours of F-16

Since the Sears-Haack slender body has continuous first derivative, the stability of wave drag computation could be achieved by using sufficient number of cross-sections. On the other hand, the geometry of F-16 aircraft has discontinuities which directly affects the area distribution. Five different size of elements representing the aircraft geometry are employed in order to obtain the mesh-independent solutions. 6.7 millions elements are created for the half aircraft in Fig. 5 [3].

The CFD analysis of F-16 is completed at Mach = 2 for comparison of the results which are shown in Table 4. There, it can be seen that difference between wave drag coefficients of Rallabhandi and of present study is 5.7% (Table 1).

Figure 6 represents the Mach contours of the $F - 16$.

6 Results

The optimal forms of the configurations are represented in this section. Nonlifting surfaces are modified during the optimization loop in order not to change the aerodynamic characteristics of the aircraft configurations. In other words, fuselage is reshaped to minimize wave drag coefficient. Mach cuts plays an important role on calculating wave drag force for an arbitrary shaped aircraft. Equation (81) represents the Mach angle which is used for calculating intercepted area distribution without using the *Mach cone approach*. In detail, the aerodynamic characteristics of an aircraft must remain unchanged during optimization. Therefore, lifting and control surfaces and related cross-sections are excluded for the optimization algorithm. Furthermore, the total volume of the aircraft is calculated by summing all parts despite exclusion of lifting and control surfaces. In other words, non-lifting surfaces are modified with respect to the objective function. Second, the intercepted cross-sectional area distribution for various Mach number is obtained by neglecting the small changes due to Mach cone method. It can be stated that non-lifting surfaces of a high-speed aircraft must be as smooth as possible due to avoid the flow separation and shock formation. For this reason, nonlifting surfaces such as fuselage does not have sharp changes which brings out the intercepted area distribution for various Mach numbers could be obtained with *Mach angle methodology* only [6, 8]:

$$\mu = \sin^{-1}\left(\frac{1}{Mach}\right) . \tag{83}$$

Despite the fact that the Mach number seems influential the optimization process, the optimal cross sectional area distribution is independent from the Mach number. Mach number only affects the intercepted area distribution only. Thus, Sears–Haack slender body has the minimum wave drag coefficient for a given volume and length. However, the intercepted area distribution and wave drag coefficient of it change with respect to the Mach number. To have the minimum value of the wave drag force coefficient for an aircraft, the change of first derivative of cross-sectional area distribution of the entire aircraft has to be minimum for a given volume and length. The methodology explained above is commonly used for high subsonic and supersonic aircraft design development. (*Feet and degree are used as length and angle units for all cases*.)

6.1 Conceptual Aircraft Design

Lifting and control surfaces are not modified during optimization in order not to alter aerodynamic characteristics of the aircraft. In addition, theoretical validation is the most important argument. Despite the fact that optimal shape of the conceptual aircraft design is not the best choice for manufacturability, theoretical aspect of the

Table 2 Wing specifications of the conceptual aircraft design

	Wing
Airfoil	NACA 63A304
Chord	2-24
Span	18.55
Sweep	45
Dihedral	0

Table 3 Tail specifications of the conceptual aircraft design

	Tail
Area	103.2
Sweep	45
Dihedral	12.3
Airfoil	4% BICONVEX
Span	0

Table 4 Fuselage specifications of the conceptual aircraft design

	Fuselage
Length	72.75
Volume	45

optimal form is satisfying. Specifications of the conceptual aircraft design are given in Tables 2, 3 and 4.

Figures 7 and 8 represent the initial and the optimal configurations of conceptual aircraft design. As seen in Fig. 8, wing and tail area distribution affect the fuselage shape to obtain the optimal area distribution. Furthermore, theoretical aspects of the optimization method provide the optimal conceptual aircraft configuration despite the fact that applicability to actual design projects requires advanced design methods. To illustrate this, optimal form of the geometry can be utilized by using wing-body concepts for high-speed UAVs.

Fuselage area distribution is modified as seen in Fig. 9. Total volume and length of the aircraft are kept constant during optimization.

6.2 *Supersonic Aircraft Geometry with GE F* $-$ **414**

More practical point of view than the theoretical approach can be obtained by employing supersonic aircraft on use for drag minimization. Since the cross-sectional area of air intakes are subtracted from entire area distribution with respect to linearized theory, Figs. 10, 11 and 12 represent the comparison between the optimal and initial form of the three dimensional supersonic aircraft configuration without air intakes. Tables 5, 6 and 7 represent the specifications of the supersonic aircraft.

Fig. 7 Initial conceptual aircraft geometry

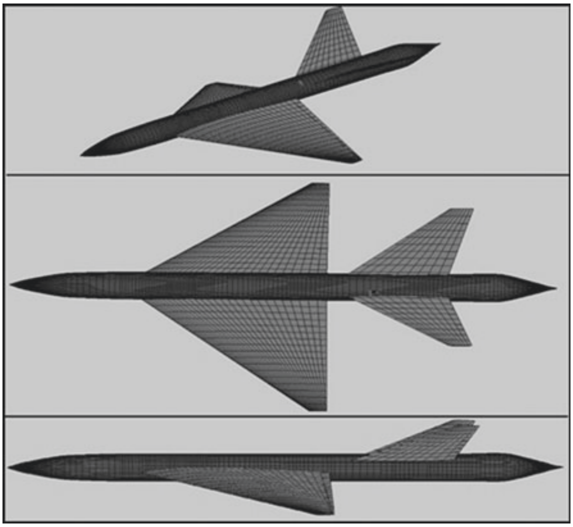
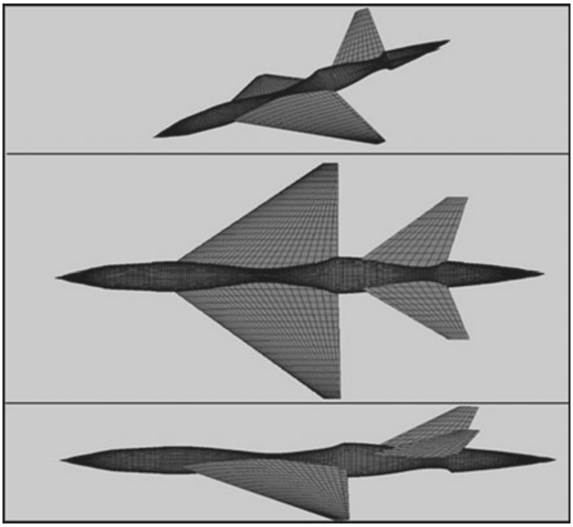


Fig. 8 Optimal conceptual aircraft geometry



The diameter and length of $GE\ F - 414$ are 3.96 ft and 15.18 ft [18]. According to these dimensions, minimum cross-sectional area for the engine region is 15.4 ft^2 (minimum cross-sectional area is calculated by multiplying the area of the engine with 1.20). Thus, the locations representing engine location are fixed to this value.

The wave drag coefficient of the supersonic aircraft is reduced from 0.185 to 0.171 with the constraints explained above. Total volume of the aircraft is not kept constant in order to avoid unnecessary increase in nose and canopy region. In detail, magnitude of areas related to engine section are fixed by employing simple calculation which is

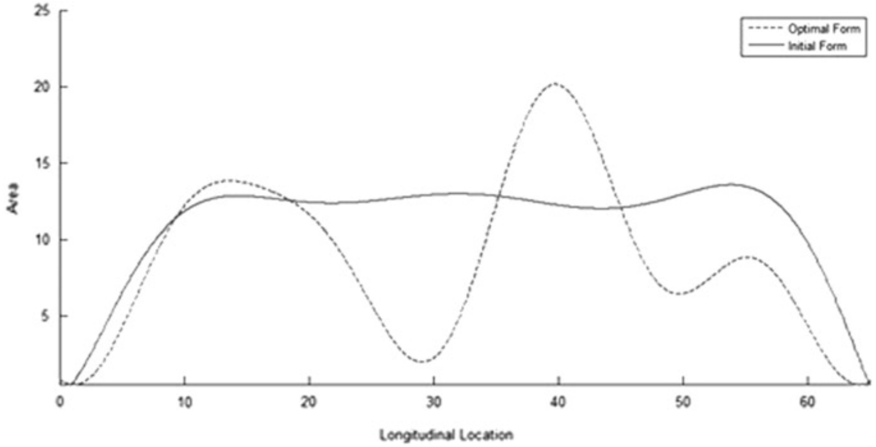


Fig. 9 Comparison of initial and optimal fuselage area distribution

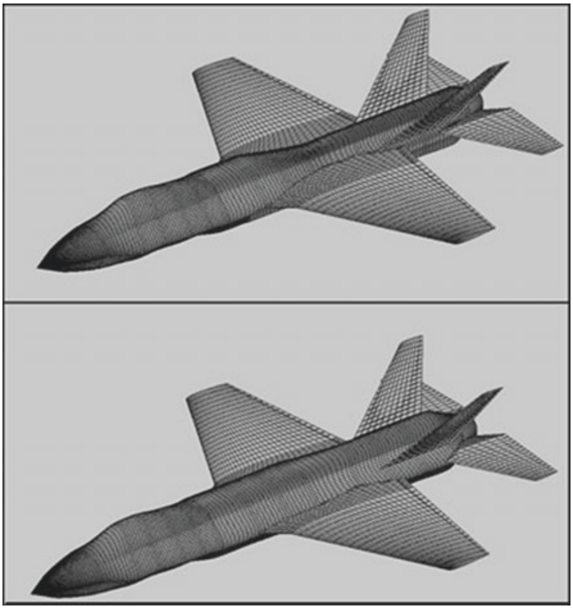


Fig. 10 Initial *(bottom)* and final *(top)* configuration of supersonic aircraft-isometric

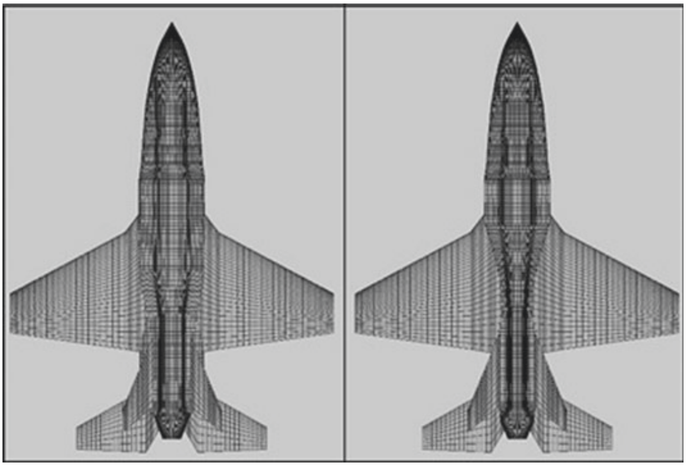


Fig. 11 Initial *(left)* and final *(right)* configuration of supersonic aircraft-top

Fig. 12 Initial *(bottom)* and final *(top)* configuration of supersonic aircraft-side

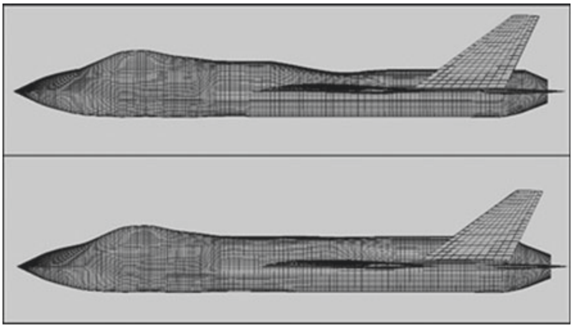


Table 5 Wing specifications of supersonic aircraft

	Section 1	Section 2
Span	5.13	12.59
Tip chord	12.73	4.26
Root chord	20.43	12.73
Sweep	52	28
Dihedral	0	0

less than the initial magnitudes [14]. Theoretically, volume participants at the front region of the aircraft increases to keep volume constant which results in impractical decision. To provide this, constraints of engine section is used for optimization. Finally, Fig. 13 represents the initial and the optimal fuselage area distribution of the supersonic aircraft geometry with $GE\ F - 414$. The lower part of the fuselage must have a place for landing gear and other components. Thus, an area reduction is applied for the upper part of the fuselage as seen in Fig. 12.

Table 6 Vertical tail specifications of supersonic aircraft

	Section 1	Section 2
Span	1.77	7.36
Tip chord	8.05	2.30
Root chord	8.05	12.73
Sweep	0	50
Dihedral	0	60

Table 7 Horizontal tail specifications of supersonic aircraft

	Section 1	Section 2
Span	2.84	7.53
Tip chord	6.99	2.37
Root chord	3.76	6.99
Sweep	29.52	29.52
Dihedral	0	60

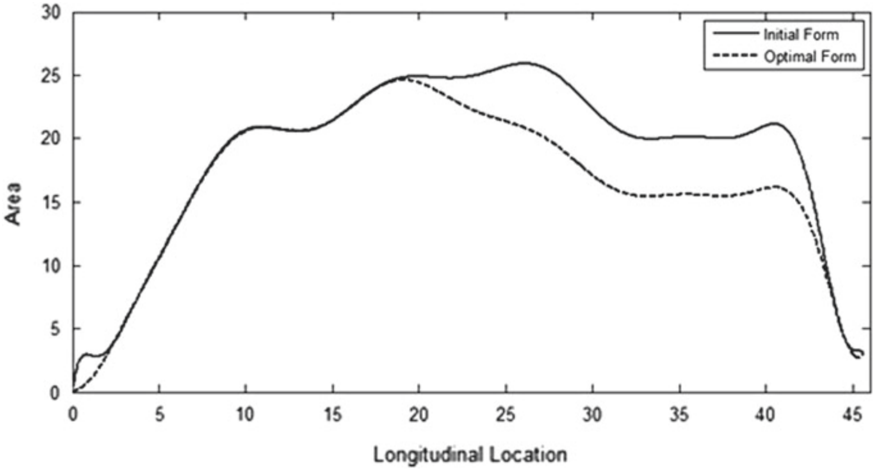


Fig. 13 Comparison of initial and optimal fuselage area distribution (supersonic aircraft configuration with $GE\ F - 414$)

7 Conclusion and Discussion

In this chapter, the numerical optimization of the wave drag is performed. At the early stages of research, a literature survey is completed on methods about wave drag calculation, and optimization. The significance of wave drag for high-speed aircraft plays major role on supersonic flow regime. Despite the fact that many other drag types play role on the calculation of the overall drag, wave drag coefficient describes the performance of aircraft at high speeds. Secondly, the solver is verified by using two different aircrafts the wave drag coefficients of which are obtained from literature. It is seen that the difference between the results of the actual study and the literature results are in sufficiently close agreement so as to implement the optimization algorithm. Results are obtained from computational fluid dynamics simulations with a variety of supersonic flow speeds. $F-16$ aircraft is analyzed and obtained that error is smaller than 8%. Next, test cases are created with respect to the aerodynamic parameters. The case matrix is generated to analyze the effect of each aerodynamic parameter such as dihedral angle and area of the control surfaces. It is verified that various types of aircrafts could be optimized by using the algorithm. Although the optimal shape of each configuration has the smallest wave drag coefficient for the given volume and length, the manufacturability of these aircraft remains vague. In addition, geometry of the aircraft on use is optimized by employing the constraints related to the engine size in order to show the algorithm can be used not only theoretical but also practical approaches. Finally, the program has the ability to optimize the entire configuration. However, parts having no effect on the aerodynamic characteristics are enforced to body shape change. A main reason behind this is preventing from additional aerodynamic trade-off analysis while generating the final configuration of the designed aircraft. In conclusion, aircrafts which are environmentally friendly by saving fuel, and provides high-level security with better performance can be obtained as a result of the study. As a future aim, additional objective functions could be added to the program. Maximization of lift will be complementary for the optimization problem of the complete aircraft post-design.

References

1. Ashley, H., Landahl, M.: Aerodynamics of Wings and Bodies. Dover Publications, New York (1965)
2. Cahn, M.S., Olstad, W.B.: A Numerical Method for Evaluating Wave Drag. National Advisory Committee for Aeronautics, Langley Field (1958)
3. Citak, C.: Wave Drag Optimization of High Speed Aircraft. Middle East Technical University, Ankara (2015)
4. Citak, C., Ozgen, S.: Sesustu Hava Araclarinin Dalga Surukleme Katsayilarinin Sayisal Yontemlerle Hesaplanmasi, SAVTEK 2014, Ankara (2014)
5. Eminton, E.: On the Numerical Evaluation of the Drag Integral. Ministry of Aviation, London (1961)
6. Entsminger, A., David, G., Will G.: General Dynamics F-16 Fighting Falcon (2014)
7. Geiselhart, K.: Integration of Multifidelity Multidisciplinary Computer Codes for Design and Analysis of Supersonic Aircraft. AIAA
8. Griva, I., Nash, S.G., Sofer, A.: Linear and Nonlinear Optimization. Society for Industrial and Applied Mathematics, Philadelphia (2008)
9. Hepperle, M.: The Sonic Cruiser - A Concept Analysis, International Symposium “Aviation Technologies of the XXI Century: New Aircraft Concepts and Flight Simulation”, Berlin, (2002)
10. Hutchison, M.G.: Multidisciplinary Optimization of High - Speed Civil Transport Configurations Using Variable - Complexity Modeling. Virginia Polytechnic Institute and State University, Blacksburg (1993)
11. Knill, O.: Multivariable Calculus - Lecture 21: Greens theorem. Harvard University, (2011)

12. Kribler, T.: A Conceptual Design Methodology to Predict the Wave Drag of a Transonic Wing, Aerodynamic Design and Optimization of Flight Vehicles in a Concurrent Multi-Disciplinary Environment, Ottawa (1999)
13. Rallabhandi, S.K., Mavris, D.N.: An Unstructured Wave Drag Code for Preliminary Design of Future Supersonic Aircraft. AIAA, Orlando (2003)
14. Raymer, Daniel P.: Aircraft Design: A Conceptual Approach. AIAA Education Series. Wiley, Reston (2012)
15. Roy Jr., V.: Harris, : An Analysis and Correlation of Aircraft Wave Drag. NASA Technical Memorandum. Langley Station, Hampton (1964)
16. Strang, W.J., McKinlay, R.: Concorde in Service. Aeronaut. J. **83**(818), 39–52 (1979)
17. Ward, G.N.: Linearized Theory of Steady High - Speed Flow. Cambridge University Press, Cambridge (1955)
18. Wikipedia (2014) Wikipedia GE F414
19. Wilhite, A.W.: An Overview of NASA's High - Speed Research Program, ICAS (2000)

Risk Modeling in Optimization Problems via Value at Risk, Conditional Value at Risk, and Its Robustification

Zeynep Cobandag Guloglu and Gerhard Wilhelm Weber

Abstract In this chapter, we explore the portfolio selection problem involving uncertainty, in other words: risk. To deal with this uncertainty, we will utilize Value at Risk (VaR) and Conditional Value at Risk (CVaR). Moreover, we present a Robust Optimization method for specifying the parameter uncertainty while minimizing the Conditional Value at Risk. We investigate optimization problems in order to minimize CVaR. Our approach consists in the use of robust optimization techniques for minimization of CVaR. We research Robust CVaR (RCVaR) optimization models under ellipsoidal uncertainty. Finally, we conclude that one can control the parameteric uncertainty with some robust distribution assumptions and obtain certain optimal solutions.

Keywords Coherency · Value at Risk · Conditional Value at Risk · Robust Conditional Value at Risk · Optimization · Robust optimization

1 Introduction

Quantifying risk in a portfolio optimization problem is an issue that should be taken seriously because it is the first step of portfolio risk management. Especially, the high volatile nature of financial markets necessitates comprehensive risk analysis and risk measurement which generate optimal solutions. *Value at Risk*, or simply *VaR*, is a specified quantile based risk measure which has been increasingly used as a risk management tool, especially, after the Risk Metrics document of Morgan [16]. VaR is so often used because it is easy to compute and understand, however, it has

Z. Cobandag Guloglu (✉)
Institute of Social Science, Istanbul Technical University (ITU), Istanbul, Turkey
e-mail: cobandag@itu.edu.tr

G.W. Weber
Institute of Applied Mathematics, Middle East Technical University (METU),
Ankara, Turkey
e-mail: gweber@metu.edu.tr

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_8

some undesirable properties which are widely criticized by researchers. For example, Uryasev [24] discussed that VaR does not take into account the risk that exceeds VaR, and for different confidence levels it can provide conflicting results. Artzner, Delbaen, Eber and Heath [2] in 1998 stated some axioms which a risk measure should satisfy. A risk measure that satisfies these axioms is called a *coherent* risk measure. After their study in 1998, coherent risk measure concept has become a criterion to evaluate risk measures. VaR is not a coherent risk measure since it fails to hold subadditivity axiom of coherence. Further, Acerbi and Tasche [1] in 2001 noted that VaR contrasts with portfolio diversification (diversification reduces risk) due to it is failure in holding axiom and they then commented that VaR is not a risk measure, since a risk measure cannot violate the subadditivity axiom. Since VaR is not coherent, risk professionals have started to search for an alternative risk measure to VaR which is coherent [1].

As a measure of downside risk, *Conditional Value at Risk,* or simply *CVaR,* came into existence and exhibited some attractive properties. CVaR is defined as the expected loss under the condition that loss exceeds VaR. First of all, CVaR is attractive since it is a coherent risk measure. Since CVaR is convex, it is relatively easy to control and optimize. Rockafellar and Uryasev [25] in 2000 first defined CVaR as a solution of an optimization problem and they have stated a minimization formula. They showed with numerical experiments, that minimization of CVaR also leads to near optimal solutions in VaR since VaR never exceeds CVaR. In their study [25], they created a new technique, minimization formula, and using this technique one can compute the VaR value and optimize CVaR at the same time. Rockafellar and Uryasev [24] in 2002 noted that as a tool in optimization modeling, CVaR has predominant properties in many respects. CVaR has a computational advantage over VaR, such as CVaR can be employed for optimizing over very large numbers of instruments and scenarios by simply using the minimization formula and applying this formula to a linear programming technique [24]. Researchers are still continuing to study on the mathematical and computational properties of CVaR [17, 18, 27].

However, the optimization processes have been recently illustrated as possibly being weak, since they lead to solutions which heavily depend on parameter relaxation. This dependence makes the theoretical and numerical results highly unreliable for practical purposes. An approach that can overcome this drawback is *robust optimization*. Robust optimization is a type of mathematical optimization problem and methodology which focuses on parameter uncertainty [3–5]. It assumes that parameters are only known to belong to certain intervals with a certain confidence level, and their value can cover certain variation ranges. By treating the uncertainty in parameters deterministically, one can have a more conservative portfolio selection. Pinar and Tutuncu [21] used robust models for risky financial contracts and they stated that the most robust profit opportunity can be solved as a convex quadratic programming problem. Further, Pinar [20] applied a robust multi-period portfolio selection problem based on minimizing one sided return from a target return level. The study found relatively stable portfolios in face of market risk and showed that robust models diminish the variability of a portfolio value. Chen et al. [29] and Ghaoui et al. [11] investigated robust portfolio selection using Worst Case Value at Risk. Chen et al. [29] provided robust Worst-Case VaR optimization under an interval random uncertainty set. With some numerical experiments they presented that the behaviour of portfolios can be improved significantly by using the robust Worst-Case VaR. Further, Ghaoui et al. [11] showed that optimizing the Worst-Case VaR can be solved exactly by solving convex, finite dimensional problems. Quaranta and Zaffaroni [31] in their study applied Robust Conditional Value at Risk methodology to deal with uncertainty in the portfolio selection problem. In their study [31], they converted the Rockafellar and Uryasev minimization model into a linear robust model. However, their study resulted with very conservative results. Zhu et al. [35] applied Worst Conditional Value at Risk Approach as an effective alternative to CVaR in complex financial markets in case, where the exit time of investors is uncertain. This makes the model interesting to risk and asset managers [35]. Zhifeng and Li in their study [33] also used robust optimization techniques to minimize CVaR of a portfolio. In their new optimization method, they captured asymmetries in the return distributions by using a robust optimization methodology. Zhu et al. [34] showed in their research paper that min-max portfolio optimization with an ellipsoidal ucertanity set is more attractable than other uncertianty set structures. They further used this min-max portfolio optimization model in CVaR robust optimization. They have stated that when the confidence level is high, CVaR robust optimization focuses on a small set of extreme mean loss scenarios and the resulting portfolios are optimal against the average of these extreme mean loss scenarios and tend to be more robust.

Some of the existing part of literature focuses on not only the robust formulation of robust portfolio selection but also the size and the shape of the set of the uncertain parameters in the robust portfolios. Here, the size of the set gives the probability that the uncertain parameter takes on a value in the set, while the shape of the set shows the robust optimization problem complexity [37]. Goldfarb and Iyengar [12] specified a factor model for the shape and the set of uncertain parameters in robust selection problems. By using this factor model framework, some new robust risk measures are proposed. The reason for the choice of factor models in robust risk measures is that the resulting problem can be formulated in a tractable way. Zhang and Chen [32] showed a new risk measure for the optimal selection with specification of a factor model. In their framework, the uncertainty in the market parameters is unknown and bounded, and optimization problems are solved assuming worst case behavior of these uncertainties. Gotoh, Shinozaki and Takeda [13] studied on the use of factor models in coherent risk minimization. In their study, they applied a simplified version to the factor model based on CVaR minimization, and showed that it improves the performance, achieving better CVaR, turnover, standard deviation and Sharpe ratio than the empirical CVaR minimization and market benchmarks.

The main objective of this chapter is to quantify the risk in an optimization problem from the view of a risk averse optimization. In Sect. 2, we shall shortly describe the coherent risk measure concept and we will present and compare the properties of VaR and CVaR both in practical and theoretical settings. With the motivation in the study [11], we shall extend the Robust VaR results to Robust CVaR and we will provide a robust optimization method for minimizing the CVaR of a portfolio. In Sect. 3, we will state applications of robust optimization methodologies which are described in the study [4] for the minimization of the conditional value at risk of a portfolio. Finally, in Sect. 4, we will conclude our results.

2 Methodology

The main objective of this study consists in modeling risk within an optimization problem from the viewpoint of a risk averse investor. Before stating the optimization problem, we will briefly review the coherent risk measure, VaR and CVaR concepts.

2.1 Coherent Risk Measures

In 1997, Artzner, Delbaen, Eber and Heath introduced the concept coherent risk measures. In their paper, they defined a complete set of axioms that have to be satisfied by a measure of risks in generalized sense [1].

Definition 1 *(Coherent Risk Measures)* Let Ω be a fixed set of scenarios. L^2 denotes the set of all functions on Ω relative to the probability measure P. Then, a risk measure ρ is a mapping from L^2 to $(-\infty, \infty]$, i.e., $\rho : L^2 \rightarrow (-\infty, \infty]$. A measure of risk ρ is called *coherent* if it satisfies the following four axioms:

1. This axiom is called as the *translation invariance* axiom of a risk measure. For all random losses X and constants α, $\rho(X + \alpha) = \rho(X) + \alpha$.
2. This axiom is called as the *subadditivity* axiom of a risk measure. For all random losses (or costs) X and Y, $\rho(X + Y) \leq \rho(X) + \rho(Y)$.
3. This axiom is called as the *positive homogeneity* axiom of a risk measure. For all $\lambda \geq 0$ and random losses X, $\rho(\lambda X) = \lambda\rho(X)$.
4. This axiom is called as the *monotonicity* axiom of a risk measure. If $X \leq Y$ for each scenario, then, $\rho(X) \leq \rho(Y)$.

Artzner, Delbaen, Eber and Heath's Coherent Measures of Risk study is important since it defined properties of portfolio statistics in order to be an appropriate risk measure for the first time. Thus, the risk management process has its own scientific rules with this deductive framework [1]. Furthermore, the theory of coherent risk measures relies on the idea that a sensible measure of risk is coherent with the finance theory and portfolio theory [9]. The coherence axioms concretize the risk measure properties in the statistics of portfolio theory [1]. One of the most important consequences of coherency for portfolio optimization is that it preserves convexity.

2.2 Value at Risk

For a fixed α-quantile, in other words: for a fixed confidence level $\alpha \in (0, 1)$, and a random variable X, the *Value at Risk (VaR)* level at α is defined as:

$$VaR_\alpha(X) = -q_\alpha^+(X) = q_{1-\alpha}^-(-X) = \inf\{\beta \mid F_X(\beta) \geq \alpha\}. \tag{1}$$

Let Ω be a fixed set of scenarios. Costs or losses of a financial position can be considered as a mapping from $X\colon \Omega \to \mathbb{R}$, where positive outcomes $X(\omega)$ of X are disliked, while the negative outcomes are liked at the end of the trading period if the scenario $\omega \in \Omega$ is realized. Furthermore, we should note that X belongs to a linear L^2 space relative to probability space P on Ω, which means $E[X^2] < 0$. VaR as a risk measure assigns to each random cost $X \in L^2$ a numerical quantity. Here, we should state that $z = f(x, y)$ represents the cost function. Moreover, x is the decision vector, $x = (x_1, x_2, ..., x_q) \in \mathbb{R}^q \in S$ where $S = \{x = (x_1, x_2, ..., x_q) \mid x_j \geq 0 (j = 1, 2, ..., q), x_1 + x_2 + ... + x_q = 1\}$, and y is a random variable on the probability space (Ω, F, P) representing the uncertainties that can affect the cost. The underlying probability distribution of y in $\mathbb{R}^q$ will be assumed to have a density denoted by $p(y)$. With a known probability distribution of y as a random variable, z will also be a random variable as like X. The distribution of z depends on the decision vector. Here, $F_X(\beta)$ is the cumulative distribution function for z. When the confidence level α is given, the probability of $f(x, y)$ not exceeding a given threshold β is shown by

$$F_X(\beta) = \int_{f(x,y)\leq\beta} p(y)dy. \tag{2}$$

The statistic $VaR_\alpha(X) = -q_\alpha^+(X)$ responds the minimum loss that can occur in the set of all α-quantiles of X over a holding period of time. Thus, VaR equals to the α-percentile of the loss distribution (α is the smallest value such that the probability that losses exceed or equal to this value is greater or equal to α). VaR is based on probabilities, so it cannot be established on certainty, but is rather a level of confidence which is selected by the user in advance. As a risk measure VaR satisfies translation invariance, positive homogeneity and monotonicity, however, it fails to hold subadditivity property [2]. Thus, VaR is not a coherent risk measure. It is known that portfolio diversification always leads to risk reduction. However, VaR contrasts with portfolio diversification [1]. In the paper [1], it is strongly believed that VaR is not a risk measure, since a risk measure can not violate the subadditivity axiom.

Moreover, VaR is not a convex risk measure. This is due to the fact that subadditivity and positive homogeneity together sufficiently show the convexity of a function, and VaR fails to satisfy subadditivity property. In an optimization problem, VaR may come out with many local minima which is a result of VaR being is not convex [22]. It should be noted that in the particular process of risk minimization, only strictly convex surfaces lead to local minima as unique globally optimal solutions [1]. Thus, in optimization problems, since VaR is non-convex, it has many extrema, and that makes it difficult to control and optimize.

Value at Risk (VaR) is one of the most widely used tools for managing risk, however, it has some undesirable properties which are widely criticized by researches, such as it is not a coherent measure of risk and it is difficult to optimize VaR when it is calculated from scenarios [25]. In addition to these two undesirable properties, VaR is a model-dependent measure of risk.

2.3 Conditional Value at Risk

Value at Risk is the predicted worst case loss of at a specified 1-α confidence level of a portfolio over a holding period of time. Differently from VaR, *Conditional Value at Risk (CVaR)* gives the expected loss that can occur in 1-α confidence of a portfolio over a holding period of time, if the portfolio distribution function is continuous. Conditional Value at Risk (CVaR) measures how much we lose on the average given we exceed our VaR. Thus, it is a measure to capture losses beyond VaR. Formally, CVaR is defined as in the following manner [24]:

$$CVaR_\alpha(X) := \phi_\alpha(X) := \text{ mean of the } \alpha\text{-tail distribution of } X. \tag{3}$$

Moreover, CVaR can be defined as the conditional expectation of the loss related to X that loss equals or is greater than $q_\alpha(X)$:

$$CVaR_\alpha(X) = \phi_\alpha(X) = \mathbb{E}\{X : X \geq q_\alpha(X)\}. \tag{4}$$

Now we focus on a vector variable x of certain realizations of the random variable X where x serves as decision variable, representing, e.g., a portfolio. Let $q_\alpha(x)$ be the VaR_α of a loss function $f(x, y)$. Then, $CVaR_\alpha(x)$ is defined as:

$$\begin{aligned} CVaR_\alpha(x) = \phi_\alpha(x) &= \mathbb{E}\{f(x, y) : f(x, y) \geq q_\alpha(x)\} \\ &= \frac{1}{1-\alpha} \int_{f(x,y) \geq q_\alpha(x)} f(x, y) p(y) dy. \end{aligned} \tag{5}$$

Here, the distribution function in Eq. (5) is defined as α-tail distribution $z = f(x, y)$ and it is truly another distribution. This new distribution function is non-decreasing and right continuous, and it is obtained by rescaling the distribution function of $z = f(x, y)$ in the interval $[\alpha, 1]$ [24].

CVaR is a coherent risk measure in the basic sense since it satisfies the properties of translation invariance, positive homogeneity and monotonicity and subadditivity [22].

Rockafellar and Uryasev [25] showed that CVaR and VaR of a loss function $z = f(x, y)$ can be computed by solving a basic, one-dimensional, convex optimization problem under a specified confidence level α, respectively. In the study [25], the main approach is to benefit from a special convex function $F_\alpha(x, \beta)$ to characterize $\phi_\alpha(x)$ and $q_\alpha(x)$. Thus, the characterization function of $\phi_\alpha(x)$ and $q_\alpha(x)$ is defined as follows [24]:

$$F_\alpha(x, \beta) := \beta + (1 - \alpha)^{-1}\mathbb{E}\{[f(x, y) - \beta]^+\},$$
$$\text{where } [f(x, y) - \alpha]^+ = [t]^+(:= \max\{0, t\}). \tag{6}$$

Theorem 1 (Optimization of CVaR) [25] *If we minimize $F_\alpha(x, \beta)$ over all $(x, \beta) \in S \times \mathbb{R}$, it gives us the equivalent result of minimizing the CVaR value $\phi_\alpha(x)$ with respect to $x \in S$ which is:*

$$\min_x \phi_\alpha(x) = \min_{x,\beta} F_\alpha(x, \beta). \tag{7}$$

In our study, we will state the optimization problem which focuses on capturing risk by CVaR. Then, optimization problem has the following form:

$$\begin{aligned} &\text{minimize}_x CVaR_\alpha(x) \\ &\quad\text{subject to } x \in S, \end{aligned} \tag{8}$$

where α represents the desired confidence level and $S = \{x = (x_1, x_2, ..., x_q) \mid x_j \geq 0(j = 1, 2, ..., q), x_1 + x_2 + ... + x_q = 1\}$. Here, x_i is the decision variable for the portfolio weight for sub-portfolio of i. Now, with the help of Theorem 1, we can convert the optimization problem in Eq. (8) to a linear programming problem as follows:

$$\begin{aligned} &\text{minimize}_{\beta,x}\ \beta + (1 - \alpha)^{-1} \int_{y \in \mathbb{R}^q} [f(x, y) - \beta]^+ p(y) dy \\ &\qquad\qquad\qquad\qquad\qquad \text{subject to } x \in S. \end{aligned} \tag{9}$$

However, the joint probability distribution of returns $p(y)$ is unknown which makes the problem in Eq. (9) still hard to solve. Before, in our study [8], we did not have any assumptions about the density $p(y)$ for simplicity, and we generated a sample from $p(y)$ only using some algorithms. But, most frequently, the decision making process is influenced by uncertain parameters, so we cannot ignore the possible implementation errors [31]. Now, to deal with data uncertainty, we shall utilize the Robust CVaR (RCVaR) approach.

2.4 Robust Conditional Value at Risk

In order to minimize implementation errors, we utilize robust estimation methods which are described in the paper Ben-Tal et al. [4]. With this method, we can control the parameter uncertainty with some steady distribution assumptions. Using this method, we can reduce modeling risk which arises due to parameter uncertainty.

In CVaR optimization problems, uncertainty is related to the distribution of portfolio return. Here, we will consider Robust CVaR in the situation, where the underlying probability distribution of return data is partially known [37].

Therefore, we will assume that the density function of portfolio return is only known to belong to a certain set P of distributions, i.e., $p(\cdot) \in P$, which covers all the possible distribution scenarios [36]. Our aim is to compute the CVaR value assuming the worst case of underlying probability distribution based on a special certain set P. Referring to the papers of Zhu and Fukushima [35], we define the *Robust CVaR (RCVaR)* for fixed $x \in S$ with respect to P as:

$$RCVaR_{\alpha}(x) := \sup_{p(\cdot)\in P} CVaR_{\alpha}(x). \tag{10}$$

Now, we shall firstly assume that y follows a discrete distribution. This assumption still contributes to the case of a continuous distribution in the CVaR formula. In fact, by sampling the probability distribution of y and its density $p(y)$, the integral of continuous distribution can be approximated [36]. Moreover, we will investigate an ellipsoidal uncertainty set which is a special case of P. We have chosen ellipsoidal uncertainty set structure since it is not only easy to specify but also tractable for practical usage [36].

Let a random variable y have a sample space which is given by $\{y_{[1]}, ..., y_{[q]}\}$ with discrete probability $Pr\{y_{[i]}\} = \pi_i$ and $\sum_{j=1}^{q} \pi_j = 1$, $\pi_i \geq 0 (i = 1, 2, ..., q)$. Further, we denote probability $\pi = (\pi_1, \pi_2, ..., \pi_q)^T$ and define:

$$H_{\alpha}(x, \beta, \pi) := \beta + \frac{1}{(1-\alpha)} \sum_{k=1}^{q} \prod_{k} \pi_k [f(x, y_k) - \beta]^{+}. \tag{11}$$

Referring to Rockafellar's and Uryasev's fundamental minimization formula (2000), the minimization of CVaR value with respect to x and π is same as minimizing the function in Eq. (11) with respect to $\beta \in \mathbb{R}$, as follows:

$$\min_x CVaR_{\alpha}(x, \pi) = \min_{\beta\in\mathbb{R}} H_{\alpha}(x, \beta, \pi). \tag{12}$$

Especially for any discrete distribution, we will present P as P_{π} which is a subset of $\mathbb{R}^q$. Then, RCVaR is defined as:

$$RCVaR_{\alpha}(x) := \sup_{\pi\in P_{\pi}} CVaR_{\alpha}(x, \pi) \tag{13}$$

or, equivalently,

$$RCVaR_{\alpha}(x) := \sup_{\pi\in P_{\pi}} \min_{\beta\in\mathbb{R}} H_{\alpha}(x, \beta, \pi). \tag{14}$$

Now, we will start to discuss computational aspects of minimization of RCVaR. We want to minimize $RCVaR_\alpha(x)$ over $x \in S$. First of all, we will consider the following optimization problem:

$$\text{minimize}_{\beta,x}\ \beta + \frac{1}{(1-\alpha)} \sum_{k=1}^{q} \pi_k [f(x, y_k) - \beta]^+ \qquad \text{subject to } x \in S. \tag{15}$$

In the linear problem of Eq. (15) we pay attention to the fact that the term $[f(x, y_k) - \beta]^+$ in the objective function be simplified. This can be done by using new variables instead of $[f(x, y_k) - \beta]^+$. First, let $k_j := f(x, y_j) - \beta$ for all $j = 1, 2, ..., q$. Then, let be given the variables $u_j = k_j^+$. Another way to state, $u_j = k_j$ if $k_j \geq 0$, and $u_j = 0$ otherwise [6]. Thus, we should change the problem of Eq. (15) according to these new variables and should add new constraints. Then, the problem in Eq. (15) is equivalent to the following linear program [25]:

$$\begin{gathered}\text{minimize}_{\beta,x,u}\ \beta + \tfrac{1}{(1-\alpha)} \sum_{k=1}^{q} \pi_k u_k \\ \text{subject to} \\ u_k + x^T y_k + \beta \geq 0,\ u_k \geq 0 (k = 1, 2, ..., q),\ x \in S.\end{gathered} \tag{16}$$

Further, one can always convert Eq. (16) into an equivalent problem in which all the complex terms are put into constraints, namely [22]:

$$\begin{gathered}\text{minimize}_{\beta,x,u,t}\ t \\ \text{subject to} \\ \beta + \tfrac{1}{(1-\alpha)} (\pi_k)^T u_k \leq t, \\ u_k + x^T y_k + \beta \geq 0,\ u_k \geq 0 (k = 1, 2, ..., q),\ x \in S.\end{gathered} \tag{17}$$

Theorem 2 [36] *If $P_\pi \in \mathbb{R}^q$ is a compact convex set, then for each x, we have [36]:*

$$RCVaR_\alpha(x) := \min_{\beta \in \mathbb{R}} \max_{\pi \in P_\pi} H_\alpha(x, \beta, \pi).$$

Theorem 2 indicates that the problem of minimizing $RCVaR_\alpha(x)$ over $x \in S$ is equivalent to following minimization problem:

$$\begin{gathered}\text{minimize}_{\beta,x,u,t}\ t \\ \text{subject to} \\ \max_{\pi \in P_\pi} \beta + \tfrac{1}{(1-\alpha)} (\pi_k)^T u_k \leq t, \\ u_k + x^T y_k + \beta \geq 0,\ u_k \geq 0 (k = 1, 2, ..., q),\ x \in S.\end{gathered} \tag{18}$$

However, the optimization problem in Eq. (18) is still not appropriate for application since it includes a max operation. Now, we will specify the uncertainty set. Let us assume that π belongs to an *ellipsoid* set P_π^E, i.e.,

$$P_{\pi}^{E} := \{\pi : \pi^0 + \mathcal{A}d, \mathbf{1}^T\mathcal{A}d = 0, \pi^0 + \mathcal{A}d \geq 0, \sqrt{d^T d} \leq 1\}, \tag{19}$$

where π^0 is a nominal distribution having the center of the ellipsoid, and $\mathcal{A} \in \mathbb{R}^{q\times q}$ is the scaling matrix of the ellipsoid. By $\mathbf{1}$ we denote the vector $(1, 1, ..., 1)^T$ in $\mathbb{R}^q$. We have the conditions $\mathbf{1}^T\mathcal{A}d = 0, \pi^0 + \mathcal{A}d \geq 0$ in order to provide that π to be a probability distribution [36]. Since

$$\beta + \frac{1}{(1-\alpha)}(\pi)^T u = \beta + \frac{1}{(1-\alpha)}(\pi^0)^T u + \frac{1}{(1-\alpha)}u^T(\mathcal{A}d), \tag{20}$$

we have

$$\max_{\pi\in P_{\pi}^{E}} \beta + \frac{1}{(1-\alpha)}(\pi)^T u = \beta + \frac{1}{(1-\alpha)}(\pi^0)^T u + \frac{\gamma(u)}{(1-\alpha)}, \tag{21}$$

where $\gamma(u)$ is the optimal value for the following linear program:

$$\begin{array}{c} \text{maximize}_{d\in\mathbb{R}^q}\ u^T(\mathcal{A}d) \\ \text{subject to} \\ \mathbf{1}^T\mathcal{A}d = 0, \pi^0 + \mathcal{A}d \geq 0, \sqrt{d^T d} \leq 1. \end{array} \tag{22}$$

Furthermore, the dual of Eq. (14) can be written as follows [36]:

$$\begin{array}{c} \text{minimize}_{(\zeta,w,v,z)\in\mathbb{R}\times\mathbb{R}^q\times\mathbb{R}^q\times\mathbb{R}}\ \zeta + (\pi^0)^T w \\ \text{subject to} \\ -v - \mathcal{A}^T w + \mathcal{A}^T\mathbf{1}z = \mathcal{A}^T u, \\ \|v\|_2 \leq \zeta, w \geq 0. \end{array} \tag{23}$$

In this case, we can equivalently reformulate the optimization problem as follows:

$$\begin{array}{c} \text{minimize}_{\beta,x,u,t}\ t \\ \text{subject to} \\ \beta + \frac{1}{(1-\alpha)}(\pi^0)^T u + \frac{1}{(1-\alpha)}(\zeta + (\pi^0)^T w) \leq t, \\ -v - \mathcal{A}^T w + \mathcal{A}^T\mathbf{1}z = \mathcal{A}^T u, \|v\|_2 \leq \zeta, w \geq 0, \\ u_k + x^T y_k + \beta \geq 0, u_k \geq 0 (k = 1, 2, ..., q), x \in S. \end{array} \tag{24}$$

3 Results and Discussion

This chapter focuses on modeling the uncertainty in optimization problems. We have stated three different risk measure called VaR, CVaR and RCVaR, respectively. In optimization problems, VaR is undesirable since it contrasts with portfolio diversification. Furthermore, VaR is difficult to optimize since it is not a convex risk measure. An alternative to VaR is CVaR, which is very sound when we compare it with VaR. Conditional Value at Risk is a coherent measure of risk and it can be easily optimized with a linear programming problem. In our earlier study [8], we have seen that CVaR optimization leads up to optimal portfolio results which are very sensitive to the inputs, where the inputs are estimated from the historical data. As a result, the optimal portfolio weights for some assets prone to be imprecise. In order to minimize implementation errors, we utilize robust estimation methods which are described in the paper Ben-Tal et al. [4]. Furthermore, we present a robust conditional value at risk optimization problem for discrete distribution cases of the uncertainty where the uncertainty set is ellipsoidal. With this method, one can control the parameter uncertainty with some robust distribution assumptions and have certain optimal solutions. Perhaps, the most important direction for future work is to apply robust optimization of CVaR for nonlinear loss/cost functions $f(x, y)$ and to robustify the unknown parameters that affect the return vector y with a reasonable specification of the uncertainty sets. In this respect, nonlinear cost functions can be eventually treated like linear functions when assuming *additive models* [14].

As we have shown, CVaR optimization can be applied in many areas in finance practically. For example, it can be used to calculate the risk of cost associated with the consideration of uncertainties or it can be used to solve the problems associated with a company such as reducing enterprise risks and to increase profit. The Optimization of Conditional Value-at-Risk work of Rockafeller and Uryasev [25] demonstrates hedging of a portfolio of options (target portfolio) through a portfolio of stocks, indices, and options (hedging portfolio) by using CVaR. Shang and Uryasev [26] used CVaR constraints in a cash flow matching problem. A cash flow matching problem optimizes a portfolio of given financial instruments (typically bonds) to match existing liabilities and assets over several time periods in the form of portfolio payments. The model is constructed on minimizing of cost subject to a CVaR constraint on matching liabilities/obligations over several time periods. In this chapter, we definitely demonstrate a robust methodology for minimization of CVaR so that many other applications of RCVaR in financial optimization and risk management can be applied. The difference between RCVaR and CVaR is that RCVaR finds a solution to the parameter uncertainities in CVaR. We naturally come up with a solution like this: CVaRs usage in practice is applicable for RCVaR. For Robust CVaR, one innovative application area is studied by Chan et al. [7]. In their study, they employ their own robust CVaR approach using radiation therapy treatment planning of breast cancer, where the uncertainty is in the patients breathing motion and the states of the system are the phases of the patients breathing cycle.

4 Conclusion

In this chapter, first we have stated coherent risk measures, VaR and CVaR concepts. Then, we have posed the optimization problem which minimizes the CVaR. Further, we have presented robust optimization method to deal with parameter uncertainties and, finally, we discussed the results. As for the future studies, we can consider an empirical portfolio selection problem and apply robust optimization of CVaR for cost functions and with respect to robustify unknown parameters that affect the return vector.

References

1. Acerbi, C., Tasche, D.: Expected shortfall: a natural coherent alternative to value at risk. Economic Notes of Banca Monte dei Paschi di Siena SpA **31**(2), 1–10 (2002)
2. Artzner, P., Delban, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Financ. **9**(3), 203–228 (1998)
3. Ben-Tal, A., Nemirovski, A.: Robust solutions of linear programming problems contaminated with uncertain data. Math. Program. **88**, 411–424 (2000)
4. Ben-Tal, A., Nemirovski, A.: Robust optimization-methodology and application. Math. Program. **92**, 453–480 (2002)
5. Bertsimas, D., Brown, D.B., Caramanis, C.: Theory and applications of robust optimization. University of Texas, Austin, USA, Tech. Rep. (2007)
6. Boduroglu, I.: Portfolio Optimization via a Surrogate Risk Measure: Conditional Fundamental Value at Risk (CFVaR). Namik Kemal University, Turkey (2010)
7. Chan, T.C.Y., Mahmoudzadeh, H., Purdie, T.G.: A robust-CVaR optimization approach with application to breast cancer therapy. Eur. J. Oper. Res. **238**, 876–885 (2014)
8. Cobandag, Z.: Risk Modeling in Optimization Problems Value at Risk and Conditional Value at Risk. The Department of Financial Mathematics, Term Project, IAM, METU (2011)
9. Eksi, Z., Comparative Study of Risk Measures, The Department of Financial Mathematics, Master Thesis, IAM, METU, 2005
10. Fabozzi, F.J., Huang, D., Zhou, G.: Robust portfolios: contributions from operations research and finance. Ann. Oper. Res. **176**, 191–220 (2010)
11. Ghaoui, L., Oks, M., Oustry, F.: Worst case value at risk and robust portfolio optimization: a conic programming approach. Oper. Res. **51**(4), 543–556 (2003)
12. Goldfarb, D., Iyengar, G.: Robust portfolio selection problems. Math. Oper. Res. **28**, 1–38 (2003)
13. Gotoh, J.Y., Shinozaki, K., Takeda, A.: Robust portfolio techniques for mitigating the fragility of CVaR minimization and generalization to coherent risk measures. Quan. Financ. **10**, 1621–1635 (2013)
14. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, Berlin (2009)
15. Karasozen, B., Pinar, M.C., Terlaky, T., Weber, G.W.: Feature cluster “Advances in continuous optimization”. Eur. J. Oper. Res. **169**, 1077–1078 (2006)
16. Morgan J.P., Reuters, Risk Metrics-Technical Document, Fourth Edition, 1996
17. Mafusalov, A., Uryasev, S.: Buffered Probability of Exceedance: Mathematical Properties and Optimization Algorithms, Research Report, University of Florida, USA (2014)
18. Pavlikov, K., Uryasev, S.: CVaR norm and applications in optimization, Optim. Lett. 1–22 (2014)
19. Pinar, M.C.: Static and dynamic var constrained portfolios with application to delegated portfolio management. Optimization **62**, 1419–1432 (2013)
20. Pinar, M.C.: Robust scenario optimization based on downside-risk measure for multi-period portfolio selection. OR Spectr. **29**, 295–309 (2007)
21. Pinar, M.C., Tutuncu, R.: Robust profit opportunities in risky financial portfolios. Oper. Res. Lett. **33**, 331–340 (2005)
22. Rockafellar, R.T.: Coherent approaches to risk in optimization under uncertainty. Tutor. Oper. Res. INFORMS **38–61**, 2007 (2007)

23. Rockafellar, R.T., Uryasev, S. The fundamental risk quadrangle in risk management. In: Optimization and Statistical Estimation Surveys in Operations Research and Management Science, vol.18 (2013)
24. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Financ. **26**, 1443–1471 (2002)
25. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**(3), 21–41 (2000)
26. Shang, D., Uryasev, S.: Cash flow matching problem with CVaR, Research Report, (2011)
27. Uryasev, S.: Buffered Probability of Exceedance and Buffered Service Level: Definitions and Properties, Research Report, University of Florida, USA (2014)
28. Uryasev, S., Optimization Using CVaR: Algorithms and Applications, Stochastic Optimization Lecture Notes 7, University of Florida, USA
29. Yang, D., Tan, S., Chen, W.: Worst-case VaR and robust portfolio optimization with interval random uncertainty set. Expert Syst. Appl. **38**, 64–70 (2011)
30. Zabarankin, M., Uryasev, S.: Statistical decision problems, selected concepts and portfolio safeguard case studies. In: Springer Optimization and Its Applications, vol. 85. Springer, Berlin (2014)
31. Zaffaroni, A., Quaranta, A.G.: Robust optimization of conditional value at risk and portfolio selection. J. Bank. Financ. **32**, 2046–2056 (2008)
32. Zhang, F., Chen, Z.: Robust Portfolio Selection With a Combined Tail Mean-variance and Factor Model. Jiaotong University, China
33. Zhifeng D., Li, D.: Robust conditional value-at-risk optimization for asymmetrically distributed asset returns. In The 8th International Conference on Optimization: Techniques and Applications(2010)
34. Zhu, L., Coleman, T.F., Yuying, L.: Min-max robust and CVaR robust mean-variance portfolios. J. Risk **11–3**, 1–31 (2009)
35. Zhu, S.S., Huang, D., Fabozzi, F.J., Fukushima, M.: Portfolio selection with uncertain exit time: a robust CVaR approach. J. Econ. Dyn. Control. Tech. Doc. **32**(2), 594–623 (2008)
36. Zhu, S.S., Fukushima, M. Worst-case conditional value at risk with application to robust portfolio management. Oper. Res. (2008)
37. Zhou, G., Huang, D., Fabozzi, F.: Robust portfolios: contributions from operations research and finance. Oper. Res. **176**, 191–220 (2010)

Zero Limit for Multi-D Conservation Laws with Nonlinear Dissipation and Dispersion

Joaquim M.C. Correia

Abstract We consider a class of nonlinear dissipative-dispersive perturbations of the scalar hyperbolic conservation law $\partial_t u + \operatorname{div} f(u) = 0$ and we study the convergence of the approximated solutions to its entropy solution. In particular, we obtain conditions under which the balance between dissipation and dispersion gives rise to the convergence.

Keywords Conservation law · Dissipative-dispersive perturbation · Entropy measure-valued solution

1 Introduction

Motivation. We study here the convergence as the parameters $\varepsilon > 0$ and δ tend to zero of smooth solutions $u^{\varepsilon,\delta}$ to the initial value problem for a class of nonlinear dissipative-dispersive equations of the form

$$\partial_t u + \operatorname{div} f(u) = \operatorname{div}\left(\varepsilon\, b_j\big(u, \nabla u\big) + \delta\, g(u) \sum_{k=1}^{d} \partial_{x_k} c_{jk}\big(g(u)\nabla u\big)\right)_{1\le j\le d} \tag{1}$$

$$u(x,0) = u_0^{\varepsilon,\delta}(x) \tag{2}$$

where the flux $f(u)$, the dissipation $b(u,\lambda)$, the dispersion $c(\lambda)$ and its coefficient $g(u)$ are given (usually nonlinear) smooth functions subject to a set of assumptions which we explicit below. We show that if the dissipation dominate the dispersion, then the solutions $u^{\varepsilon,\delta}$ converges strongly to the entropy weak solution of the zero

J.M.C. Correia (✉)
DMat, ECT & CIMA, IIFA, Universidade de Évora, Rua Romão Ramalho, 59,
7000-671 Évora, Portugal
e-mail: jmcorreia@uevora.pt

J.M.C. Correia
CAMGSD-LARSyS, IST, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_9

limit initial value problem with first order hyperbolic conservation law

$$\partial_t u + \operatorname{div} f(u) = 0, \qquad (x,t) \in \mathbf{R}^d \times [0, +\infty[\,, \tag{3}$$

$$u(x,0) = u_0(x), \qquad x \in \mathbf{R}^d, \tag{4}$$

where the datum u_0 is suitably approximated by the smooth data $u_0^{\varepsilon,\delta}$ without loss of generality.

The pioneer work in that subject was given by Schonbek [6, 1982] and concerned the (generalised) Korteweg-de Vries-Burgers equation

$$\partial_t u + \partial_x f(u) = \varepsilon\, \partial_x^2 u - \delta\, \partial_x^3 u.$$

In the case of a convex flux function $f(u)$, she proved the convergence under the condition that $\delta = o(\varepsilon^2)$, while the sharp condition should be $\delta = o(\varepsilon^1)$[1] according to Perthame-Ryzhic [5, 2007]. LeFloch-Natalini [4, 1997] considered the equation with nonlinear viscosity, linear capillarity and general flux function

$$\partial_t u + \partial_x f(u) = \varepsilon\, \partial_x \beta(\partial_x u) - \delta\, \partial_x^3 u$$

and proved the convergence to the entropy weak solution using the setting of DiPerna's measure-valued solutions. Then, Correia-LeFloch [1, 1998] improved the estimates in Schonbek [6] and LeFloch-Natalini [4] and treated for the first time a multidimensional equation

$$\partial_t u + \operatorname{div} f(u) = \operatorname{div} \Big(\varepsilon\, b_j(\nabla u) + \delta\, \partial_{x_j}^2 u\Big)_{1 \le j \le d}$$

with general flux function $f(u)$, nonlinear viscosity function $b(\nabla u)$ and a diagonal linear capillarity function. Here we study the case with nonlinear viscosity function $b(u, \nabla u)$ and in general nonlinear dispersion $g(u)c(g(u)\nabla u)$, but which includes the general linear case.

In all these equations, when $\delta = 0$ we reduce to the (generalised) Burgers' equations and the approximate solutions $u^{\varepsilon,0}$ converge to the solution of the (generalised) inviscid Burgers' equation, this is the *vanishing viscosity method*, see, e.g., Whitham [8] or Kružkov [2] (in the present work, it is a sub product of the convergence proof taking $\delta = 0$).

On the other hand, when $\varepsilon = 0$ and in dimension one, if we consider the flux function $f(u) = u^2$ and the linear dispersion δu_{xxx} we obtain the Korteweg-de Vries equation. The approximate solutions $u^{0,\delta}$ do not converge in a strong topology, Lax-Levermore [3, 1983]. In general such a behaviour is expected to those pure ($\varepsilon = 0$) dispersive equations.

[1] Our results do not apply to simultaneously linear dissipation and linear dispersion as the condition below $r \geq \rho + 1 + \vartheta$ implies, but it is equivalent to $\frac{\rho+2}{r+1-\vartheta} \leq 1$ and our results match for $\delta = o(\varepsilon^{\frac{\rho+2}{r+1-\vartheta}})$.

So, as parameters ε and δ vanish, we are concerned with singular limits and to ensure convergence it is necessary a dominant dissipation regime. See also the analogy between the singular limit for the Korteweg-de Vries-Burgers equation and the hydrodynamic limit of the kinetic Boltzmann equation for a rarefied gas to the continuum Euler equations of compressible gas dynamics as the Knudsen number approaches zero in "From Boltzmann to Euler: Hilbert's 6th problem revisited", Slemrod [7, 2013].

Assumptions. Now we collect for the sake of clarity all the assumptions we made. Let $u^{\varepsilon,\delta} : \mathbf{R}^d \times [0, T] \to \mathbf{R}$ be smooth solutions to the initial value problem (1)–(2) defined on the interval $[0, T]$ independent of ε, δ and decaying rapidly at infinity; $u_0^{\varepsilon,\delta}$ is a suitable approximation of $u_0 : \mathbf{R}^d \to \mathbf{R}$ the datum in (4). Throughout, it is assumed $u_0 \in L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ and the $u_0^{\varepsilon,\delta}$ are smooth functions with compact support and uniformly bounded in $L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$. Since our attention is restricted to the dissipation dominant regime we regard $\delta = \delta(\varepsilon)$ and we suppose, for simplicity, that $u_0^{\varepsilon,\delta}$ approaches u_0 in the sense that

$$\begin{aligned} &\lim_{\varepsilon\to 0+} u_0^{\varepsilon,\delta} = u_0 \quad \text{in} \;\; L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d), \\ &\|u_0^{\varepsilon,\delta}\|_{L^1\cap L^q} \le \|u_0\|_{L^1\cap L^q}. \end{aligned} \tag{5}$$

The nonlinear flux function $f : \mathbf{R} \to \mathbf{R}^d$ is continuously differentiable such that

(A_1) for some $m > 1$, $\left|f'(u)\right| = \mathscr{O}\left(|u|^{m-1}\right)$ as $|u| \to \infty$.

The continuous dissipation function $b : \mathbf{R} \times \mathbf{R}^d \to \mathbf{R}^d$ is dissipative, $\forall u \in \mathbf{R}, \lambda \in \mathbf{R}^d$, $\lambda \cdot b(u, \lambda) \ge 0$, such that

(A_2) for some $\mu \ge 0, r > 0$, $|b(u, \lambda)| = \mathscr{O}\left(|u|^{\mu}\right) \mathscr{O}\left(|\lambda|^r\right)$ as $|u|, |\lambda| \to \infty$,

which we use in the convergence proof. Also, to obtain the a priori estimates we need to establish, both, the existence of the Young measure solution and that the dissipation dominates the dispersion, then we ask the dissipation to verify

(A_3) for some $\varphi \ge 0, \vartheta < r, D > 0$, $\lambda \cdot b(u, \lambda) \ge D\,|u|^{\mu\varphi}\,|\lambda|^{r+1-\vartheta}$, $\forall u \in \mathbf{R}$, $\lambda \in \mathbf{R}^d$.

So, because of (A_2)-(A_3) compatibility, $\varphi \le 1$ and $\vartheta \ge 0$. And, supposing $u_0 \in L^q(\mathbf{R}^d)$, to ensure convergence arguments we will need also that $\mu\left(1 + r\frac{1-\varphi}{1-\vartheta}\right) \le q$.

The dispersion matrix $[c_{jk}]$ is the jacobian matrix of a continuously differentiable function $C : \mathbf{R}^d \to \mathbf{R}^d$ such that $C(0) = 0$ and

(A_4) for some $\rho > 0$, $\left\|[c_{jk}(\lambda)]\right\| = \mathscr{O}(|\lambda|^{\rho})$ as $|\lambda| \to \infty$.

Finally, the coefficient $g : \mathbf{R} \to \mathbf{R}$ is a continuously differentiable function such that

(A_5) for some $q_0 \ge 1$, $\left|g'(u)\right| = \mathscr{O}(|u|^{q_0-1})$ as $|u| \to \infty$.

The following versions of (A_4) and (A_5) will be also in use:

$(\tilde{A}_4)$ for some $\rho > 0, K_c > 0$: $\forall \lambda \in \mathbf{R}^d$ $\left\|[c_{jk}(\lambda)]\right\| \le K_c\,|\lambda|^{\rho}$;

$(\tilde{A}_5)$ for $q_0 = 0$ or some $q_0 \geq 1$, $K_g > 0$: $\forall u \in \mathbf{R} \quad |g(u)| \leq K_g\, |u|^{q_0}$.

Main Results. The main results of the present work establishes that, under rather broad assumptions, the solutions of (1)–(2) tends to the entropy weak solution of (3)–(4). And, the appropriate balance between dissipation and dispersion is given by the balance of strengths between ε and δ and of growths of b and c:

$$\delta = o(\varepsilon^{\frac{\rho+2}{r+1-\vartheta}}), \qquad r \geq \rho + 1 + \vartheta.$$

We will use the L^p-Young measure framework as given by Tartar-Schonbek-DiPerna-Szepessy, see, e.g., Correia-LeFloch [1]. So, we remember now the definition of entropy measure-valued solution and the two results we will apply here.

Definition 1 Assume that $f \in \mathrm{C}(\mathbf{R})^d$ satisfies the growth condition

$$|f(u)| = \mathscr{O}(|u|^m) \quad \text{as } |u| \to \infty, \qquad \text{for some } m \in [0, q), \tag{6}$$

and $u_0 \in L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$. A Young measure ν associated with $\{u_n\}_{n\in\mathbf{N}}$, a bounded sequence in $L^\infty((0,T); L^q(\mathbf{R}^d))$, is called an entropy measure-valued (e.m.-v.) solution to (3)–(4) if

$$\partial_t \langle \nu_{(\cdot)}, |u-k|\rangle + \operatorname{div}\langle \nu_{(\cdot)}, \operatorname{sgn}(u-k)(f(u)-f(k))\rangle \leq 0, \quad \text{for all } k \in \mathbf{R}, \tag{7}$$

in the sense of distributions on $\mathbf{R}^d \times (0,T)$ and, for all compact set $K \subseteq \mathbf{R}^d$,

$$\lim_{t\to 0^+} \frac{1}{t} \int_0^t \int_K \langle \nu_{(x,s)}, |u - u_0(x)|\,\rangle\, dxds = 0. \tag{8}$$

Lemma 1 *Let $\{u_n\}_{n\in\mathbf{N}}$ be a bounded sequence in $L^\infty((0,T); L^q(\mathbf{R}^d))$. Then there exists a subsequence denoted by $\{\tilde{u}_n\}_{n\in\mathbf{N}}$ and a weakly-$\star$ measurable mapping $\nu : \mathbf{R}^d \times (0,T) \to Prob(\mathbf{R})$ such that, for all functions $g \in C(\mathbf{R})$ satisfying*

$$g(u) = \mathscr{O}(|u|^s) \quad \textit{as } |u| \to \infty, \qquad \textit{for some } s \in [0,q), \tag{9}$$

$\langle \nu_{(x,t)}, g\rangle$ belongs to $L^\infty((0,T); L^{q/s}_{loc}(\mathbf{R}^d))$ and the following limit representation holds

$$\lim_{n\to\infty} \iint_{\mathbf{R}^d\times(0,T)} g(\tilde{u}_n(x,t))\, \phi(x,t)\; dxdt \tag{10}$$
$$= \iint_{\mathbf{R}^d\times(0,T)} \langle \nu_{(x,t)}, g\rangle\; \phi(x,t)\; dxdt$$

for all $\phi \in L^1(\mathbf{R}^d \times (0,T)) \cap L^\infty(\mathbf{R}^d \times (0,T))$.

Conversely, given ν, there exists a sequence $\{u_n\}_{n\in\mathbf{N}}$ satisfying the same conditions as above and such that (10) holds for any g satisfying (9).

The notation Prob($\mathbf{R}$) is for the space of probability measures (non-negative measures with unit total mass). Finally, the following theorem is the convergence tool we will use.

Theorem 1 *Assume that f satisfies (6) and $u_0 \in L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ for $q > 1$. Let $\{u_n\}_{n\in\mathbf{N}}$ be a bounded sequence in $L^\infty((0,T); L^q(\mathbf{R}^d))$ with associated Young measure ν. If ν is an e.m.-v. solution to (3)–(4), then*

$$\lim_{n\to\infty} u_n = u \quad in\ L^s((0,T); L^p_{loc}(\mathbf{R}^d)), \quad \forall s < \infty,\ p \in [1,q),$$

$u \in L^\infty((0,T); L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d))$ is the unique entropy solution to (3)–(4).

There are our main results[2] :

Theorem 2 *Consider the Cauchy problem (3)–(4) where the flux function f satisfies (A_1) with $m < q$ and the initial data $u_0 \in L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$. Let $u^{\varepsilon,\delta}$ be the solutions of the perturbed problem (1)–(2) where $\varepsilon > 0$ and the dissipation and the dispersion functions satisfy assumptions (A_2)-(A_3) and $(\tilde{A}_4)$-$(\tilde{A}_5)$ subject to*

$$\mu\left(1 + r\frac{1-\varphi}{1-\vartheta}\right) \le q \qquad and \qquad r \ge \rho + 1 + \vartheta,$$

$$\mu\varphi + (3 - (\rho+1)q_0)\,\frac{r+1-\vartheta}{\rho+2} \le q \qquad and \qquad \frac{\mu\varphi - 2}{r+1-\vartheta} \ge \frac{(\rho+1)q_0 - 3}{\rho+2}.$$

Then, with $\delta = o(\varepsilon^{\frac{\rho+2}{r+1-\vartheta}})$, the sequence $u^{\varepsilon,\delta}$ converges in $L^s\left((0,T); L^p_{loc}(\mathbf{R}^d)\right)$ for all $s < \infty$ and $p < q$ to a function $u \in L^\infty\left((0,T); L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)\right)$ which is the unique entropy solution to (3)–(4).

Theorem 3 *Consider the Cauchy problem (3)–(4) where the flux function f satisfies (A_1) with $m < q$ and the initial data $u_0 \in L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$. Let $u^{\varepsilon,\delta}$ be the solutions of the perturbed problem (1)–(2) where $\varepsilon > 0$ and the dissipation and the dispersion functions satisfy assumptions (A_2)-(A_3) and (A_4)-(A_5) subject to*

$$\mu\left(1 + r\frac{1-\varphi}{1-\vartheta}\right) \le q \qquad and \qquad r \ge \rho + 1 + \vartheta,$$

if $(\rho+1)q_0 \le 3$:

$$\mu\varphi + 3\,\frac{r+1-\vartheta}{2} \le q \qquad and \qquad \frac{\mu\varphi - 2}{r+1-\vartheta} \ge \frac{(\rho+1)q_0 - 3}{\rho+2},$$

if $(\rho+1)q_0 \ge 3$:

[2] They correspond, respectively, to the Propositions 4 and 3. Using the Propositions 5 and 6 we can state two more theorems, relative to the assumptions made in these propositions.

$$\frac{\mu\varphi - 2}{r + 1 - \vartheta} \geq \frac{(\rho+1)q_0 - 3}{2}.$$

Then, with $\delta = o(\varepsilon^{\frac{\rho+2}{r+1-\vartheta}})$ *the sequence* $u^{\varepsilon,\delta}$ *converges in* $L^s\left((0,T); L^p_{loc}(\mathbf{R}^d)\right)$ *for all* $s < \infty$ *and* $p < q$ *to a function* $u \in L^\infty\left((0,T); L^1(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)\right)$ *which is the unique entropy solution to* (3)–(4).

Note that we can handle arbitrary large values of q: the natural one as given by the initial data $u_0 \in L^q(\mathbf{R}^d)$. And, no more a function of the growth m of the flux function f as in Schonbek [6] and LeFloch-Natalini [4].

2 A Priori Estimates

Preliminaries. Consider the problem with general ε-dissipative and δ-dispersive terms where from now on we omit the superscripts ε, δ in the notations $u^{\varepsilon,\delta}$, $u_0^{\varepsilon,\delta}$

$$\partial_t u + \operatorname{div} f(u) = \varepsilon \operatorname{div}(\mathscr{B}) + \delta \operatorname{div}(\mathscr{C}), \qquad u(x,0) = u_0(x). \tag{11}$$

With sufficient formal conditions, we can profit of its divergence form in order to obtain (approximate) conservation laws, e.g.,

$$\frac{d}{dt}\int_{\mathbf{R}^d} u(t)\, dx = 0.$$

Let's do it. If $\eta : \mathbf{R} \to \mathbf{R}$ is a sufficiently smooth function, define $q : \mathbf{R} \to \mathbf{R}^d$ by $q_j' = \eta' f_j'$ for $j = 1, \dots, d$, then multiply (11) by $\eta'(u)$. We reach

$$\begin{aligned}\partial_t \eta(u) + \operatorname{div} q(u) = \varepsilon \operatorname{div}\big(\eta'(u)\,\mathscr{B}\big) &- \varepsilon\, \eta''(u)\, \nabla u \cdot \mathscr{B} \\ + \delta \operatorname{div}\big(\eta'(u)\,\mathscr{C}\big) &- \delta\, \eta''(u)\, \nabla u \cdot \mathscr{C}.\end{aligned} \tag{12}$$

When η is convex the ε-term containing $\eta''(u)$ has a favorable sign if $\nabla u \cdot \mathscr{B} \geq 0$, the term $\mathscr{B}$ dissipates the entropy η, and three of the remaining terms are conservative (assume that u together with their space-derivatives vanish at infinity). Integrate over $\mathbf{R}^d \times [0,t]$:

$$\int_{\mathbf{R}^d} \eta\big(u(t)\big) - \eta\,(u_0)\; dx = -\int_0^t \int_{\mathbf{R}^d} \eta''(u)\, \nabla u \cdot (\varepsilon\, \mathscr{B} + \delta\, \mathscr{C})\; dx ds.$$

Take $\eta(u) = |u|^{\alpha+1}$, so $\eta'(u) = (\alpha+1)\mathrm{sgn}(u)|u|^\alpha$, $\eta''(u) = (\alpha+1)\alpha|u|^{\alpha-1}$.

Lemma 2 *Let* $\alpha \geq 1$ *be any real and suppose that* $u_0 \in L^{\alpha+1}\big(\mathbf{R}^d\big)$, *then the solutions of* (11) *satisfy for* $t \geq 0$

$$\begin{aligned}\int_{\mathbf{R}^d} |u(t)|^{\alpha+1}\, dx = \int_{\mathbf{R}^d} |u_0|^{\alpha+1}\, dx \\ -(\alpha+1)\,\alpha \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot (\varepsilon\,\mathscr{B} + \delta\,\mathscr{C})\; dxds.\end{aligned}$$

Corollary 1 *For $\varepsilon \geq 0$, $u_0 \in L^{q_1}\big(\mathbf{R}^d\big) \cap L^{q_2}\big(\mathbf{R}^d\big)$ with $1 \leq q_1 \leq \alpha + 1 \leq q_2 \leq \infty$ and $\alpha \geq 1$, any $\mathscr{B}$-dissipative solution of (11), i.e. $\nabla u \cdot \mathscr{B} \geq 0$, verifies for $t \geq 0$*

$$\delta \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot \mathscr{C}\, dxds \leq K$$

where $K = 2^{-1} \max\Big\{1,\, \|u_0\|^{q_2}_{L^{q_1}(\mathbf{R}^d)},\, \|u_0\|^{q_2}_{L^{q_2}(\mathbf{R}^d)}\Big\}$.

If also $\delta\, \nabla u \cdot \mathscr{C} \geq 0$, then

$$\begin{aligned}&\|u(t)\|_{L^{\alpha+1}(\mathbf{R}^d)} \leq \max\Big\{\|u_0\|_{L^{q_1}(\mathbf{R}^d)},\, \|u_0\|_{L^{q_2}(\mathbf{R}^d)}\Big\};\\ &\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot \mathscr{B}\, dxds \leq K;\\ &|\,\delta\,| \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\; |\,\nabla u \cdot \mathscr{C}\,|\; dxds \leq K.\end{aligned}$$

Our purpose is to bound the $u^{\varepsilon,\delta}$ in $L^\infty\big((0,T);\, L^q\big(\mathbf{R}^d\big)\big)$. As the $\delta\, \nabla u \cdot \mathscr{C} \geq 0$ assumption is an unreasonable one in general, instead, we need to balance the dispersion with the dissipation. In [1] the $\mathscr{C} = \big(\partial^2_{x_1} u, \ldots, \partial^2_{x_d} u\big)$, therefore $\nabla u \cdot \mathscr{C} = \sum \partial_{x_j} u\, \partial^2_{x_j} u$ has no sign. But, even in the more general case, yet linear, of

$$\mathscr{C} = \left(\sum_{l=1}^{d} a_{1l}\, \partial^2_{x_1 x_l} u, \ldots, \sum_{l=1}^{d} a_{dl}\, \partial^2_{x_d x_l} u\right)$$

it provides an almost conservative structure since

$$\nabla u \cdot \mathscr{C} = 2^{-1} \sum_{j,l=1}^{d} a_{jl}\, \partial_{x_l}\, \big(\partial_{x_j} u\big)^2\,.$$

Returning back to Lemma 2, with $\alpha = 1$, we obtain the first energy estimates, which we resume in the

Proposition 1 *If $u_0 \in L^2\big(\mathbf{R}^d\big)$ and $c = \big(c_j\big)_{1 \leq j \leq d} : \mathbf{R}^d \to \mathbf{R}^d$ is a linear function with $[a_{jl}]$ matrix, then any solution of the Cauchy problems*

$$\partial_t u + div\, f(u) = \varepsilon\; div\,(\mathscr{B}) + \delta \sum_j \partial^2_{x_j} c_j(\nabla u), \qquad u(x,0) = u_0(x),$$

satisfies for $t \geq 0$

$$\int_{\mathbf{R}^d} u(t)^2\,dx + 2\,\varepsilon \int_0^t \int_{\mathbf{R}^d} \nabla u \cdot \mathscr{B}\,dx ds = \int_{\mathbf{R}^d} u_0^2\,dx$$

and assuming that $\varepsilon \geq 0$ and the solution is $\mathscr{B}$-dissipative $(\nabla u \cdot \mathscr{B} \geq 0)$

$$\|u(t)\|_{L^2(\mathbf{R}^d)} \leq \|u_0\|_{L^2(\mathbf{R}^d)}.$$

Moreover the assumption (A_3) *implies also*

$$\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi} |\nabla u|^{r+1-\vartheta}\,dx ds \leq (2D)^{-1} \|u_0\|^2_{L^2(\mathbf{R}^d)}.$$

Note that this uniform $L^2\left(\mathbf{R}^d\right)$ bound is not enough, according to (6), to treat the case of the quadratic fluxes f, such as the one in the Burgers equation or in the Korteweg-de Vries equation. Meanwhile, with $\alpha > 1$ the δ-integrand is no more conservative and to derive higher L^q energy estimates, we will see, it is a matter of competitiveness between the ε, δ strengths and the u and ∇u growths in the dissipation and in the dispersion functions.

Next, we will study a class of equations which includes the case of the full general linear dispersion

$$\operatorname{div}(\mathscr{C}) = \sum_{j,k,l} a_{jlk}\, \partial^3_{x_j x_k x_l} u,$$

in fact that with any (linear or nonlinear) dispersion of the form

$$\operatorname{div}(\mathscr{C}) = \sum_{j,k} \partial_{x_j}\big(g(u)\, \partial_{x_k} c_{jk}\big(g(u)\nabla u\big)\big)$$

where the matrix $[c_{jk}]$ is a jacobian matrix[3]. Trivial nonlinear examples are those of the form $c_{jk}\big(\nabla u\big) = c_{jk}\big(\partial_{x_j} u\big)$.

Energy estimates. We are concerned here with a priori estimates to the solutions of the initial value problem for the equations

$$\partial_t u + \operatorname{div} f(u) = \operatorname{div}\left(\varepsilon\, b_j(u, \nabla u) + \delta\, g(u) \sum_k \partial_{x_k} c_{jk}\big(g(u)\nabla u\big)\right)_{1\leq j\leq d} \qquad (13)$$

where $c_{jk} = \partial_k C_j$ $(j, k = 1, \ldots, d)$, i.e., the matrix $[c_{jk}]$ is the jacobian matrix of some function $C = \big(C_j\big)_{1\leq j\leq d} : \mathbf{R}^d \to \mathbf{R}^d$. In that case, with the notation $G'(u) =$

[3] In the linear case the necessary condition on the potential, $a_{jlk} = a_{jkl}$ $(j, l, k = 1, \ldots, d)$, is assumed done as the usual symmetrization $\tilde{a}_{jlk} = \tilde{a}_{jkl} = \frac{a_{jlk} + a_{jkl}}{2}$ keeps the equation unchanged.

$g(u)$, formula (12) rewrites as

$$\begin{aligned}
\partial_t \eta(u) + \operatorname{div} q(u) = \varepsilon \operatorname{div}\big(\eta'(u)\; b(u, \nabla u)\big) - \varepsilon\, \eta''(u)\, \nabla u \cdot b(u, \nabla u) \\
+\, \delta \sum_{j,k} \partial_{x_j}\big(\eta'(u)\, g(u)\, \partial_{x_k} c_{jk}\, (\nabla G(u))\big) \\
-\, \delta \sum_{j,k} \eta''(u)\, \partial_{x_k}\big(\partial_{x_j} G(u)\;\, c_{jk}(\nabla G(u))\big) \\
+\, \delta \sum_{j} \eta''(u)\, \partial_{x_j} C_j(\nabla G(u))
\end{aligned} \tag{14}$$

where the δ-lines above takes also on the form

$$\begin{aligned}
&+ \delta \sum_{j,k} \partial^2_{x_j x_k}\big(\eta'(u)\, g(u)\, c_{jk}(\nabla G(u))\big) \\
&- \delta \text{ div } \big(\eta'(u) \left[c_{jk}(\nabla G(u))\right] \nabla g(u)\big) \\
&+ \delta \text{ div } \big(\eta''(u)\big(C\, (\nabla G(u)) - \left[c_{jk}(\nabla G(u)) + c^T_{jk}(\nabla G(u))\right]\nabla G(u)\big)\big) \\
&+ \delta\, \eta'''(u) \nabla u \cdot \big(\left[c_{jk}(\nabla G(u))\, \right]\nabla G(u) - C\, (\nabla G(u))\big)\, .
\end{aligned} \tag{15}$$

Then, from (14) to (15), Lemma 2 becomes

Lemma 3 *Let $\alpha \geq 1$ be any real, $u_0 \in L^{\alpha+1}\left(\mathbf{R}^d\right)$, $G \in C^2\,(\mathbf{R})$ such that $G'(u) = g(u)$ and $C = \left(C_j\right)_{1 \leq j \leq d}$ be a potential of the dispersion-matrix $[c_{jk}]$, then any solution of equation (13) satisfies for $t \geq 0$*

$$\begin{aligned}
&\int_{\mathbf{R}^d} |u(t)|^{\alpha+1}\, dx + (\alpha + 1)\, \alpha\, \varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot b(u, \nabla u)\, dx ds \\
&= \int_{\mathbf{R}^d} |u_0|^{\alpha+1}\, dx + (\alpha + 1)\, \alpha\, \delta \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, div\, (\, C\, (\nabla G(u)) \\
&\qquad\qquad - \left[c^T_{jk}(\nabla G(u))\right] \nabla G(u)\big)\; dx ds.
\end{aligned} \tag{16}$$

For $\alpha \geq 2$, the δ-term also equals

$$\begin{aligned}
&(\alpha + 1)\, \alpha\, (\alpha - 1)\, \delta \int_0^t \int_{\mathbf{R}^d} sgn(u)\, |u|^{\alpha-2} \nabla u\, \cdot \\
&\qquad \big(\left[c_{jk}(\nabla G(u))\right] \nabla G(u) - C\, (\nabla G(u))\, \big)\, dx ds.
\end{aligned} \tag{17}$$

Once more with $\alpha = 1$ we obtain from (16) the first energy estimates:

Proposition 2 *Any solution of (13) with initial data $u_0 \in L^2\left(\mathbf{R}^d\right)$, dispersion-matrix $[c_{jk}]$ having a potential C and $g(u) = G'(u)$ verifies for all $t \geq 0$*

$$\int_{\mathbf{R}^d} u^2(t)\, dx + 2\, \varepsilon \int_0^t \int_{\mathbf{R}^d} \nabla u \cdot b(u, \nabla u)\, ds dx = \int_{\mathbf{R}^d} u_0^2\, dx\,, \tag{18}$$

assuming that $\varepsilon \geq 0$ *and* $b(u,\lambda)$ *is dissipative* $\big(\lambda \cdot b(u,\lambda) \geq 0\big)$ *then*

$$\|u(t)\|_{L^2(\mathbf{R}^d)} \leq \|u_0\|_{L^2(\mathbf{R}^d)}. \tag{19}$$

Moreover, if strongly we assume (A_3) *we have also that*

$$\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi} |\nabla u|^{r+1-\vartheta}\, dxds \leq (2\,D)^{-1}\, \|u_0\|^2_{L^2(\mathbf{R}^d)}. \tag{20}$$

Now, to derive higher L^p a priori estimates we use (16)–(17) in Lemma 3. We take the lower bound of the ε-term using (A_3) and the upper bound of the δ-term by (A_4)-(A_5):

$$\begin{aligned} &D\,\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi+\alpha-1} |\nabla u|^{r+1-\vartheta}\, dxds \\ &\qquad\leq \varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot\, b(u,\nabla u)\, dxds\,, \\ &|\delta| \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-2}\, |\nabla u| \,\big(\, \|[c_{jk}(\nabla G(u))]\|\, |g(u)|\, |\nabla u| + |C(\nabla G(u))|\,\big)\, dxds \\ &\qquad\leq const\, |\delta| \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-2}\big(1+|u|^{(\rho+1)q_0}\big)\, |\nabla u|^2\, \big(1+|\nabla u|^{\rho}\big)\, dxds\,. \end{aligned}$$

From (16)–(17) we deduce that

$$\begin{aligned} &\int_{\mathbf{R}^d} |u(t)|^{\alpha+1}\, dx + D\,(\alpha+1)\,\alpha\,\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi+\alpha-1} |\nabla u|^{r+1-\vartheta}\, dxds \\ &\quad\leq \int_{\mathbf{R}^d} |u(t)|^{\alpha+1}\, dx + (\alpha+1)\,\alpha\,\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-1}\, \nabla u \cdot\, b(u,\nabla u)\, dxds \\ &\quad\leq \|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + const\,(\alpha+1)\,\alpha\,(\alpha-1)\,|\delta| \\ &\qquad\qquad \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-2}\,\big(1+|u|^{(\rho+1)q_0}\big)\, |\nabla u|^2\, \big(1+|\nabla u|^{\rho}\big)\, dxds\,, \end{aligned} \tag{21}$$

consider the two inequalities we obtain with the single of each left terms, integrate over $[0,t]$ the first one and multiply by t the second, finally re-add:

$$\begin{aligned} &\int_0^t \int_{\mathbf{R}^d} |u|^{\alpha+1}\, dxds + D\,(\alpha+1)\,\alpha\, t\,\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi+\alpha-1} |\nabla u|^{r+1-\vartheta}\, dxds \\ &\quad\leq 2\,t\, \|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + 2\,t\,const\,(\alpha+1)\,\alpha\,(\alpha-1)\,|\delta| \\ &\qquad\qquad \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-2}\,\big(1+|u|^{(\rho+1)q_0}\big)\, |\nabla u|^2\, \big(1+|\nabla u|^{\rho}\big)\, dxds\,. \end{aligned} \tag{22}$$

Now we apply Young's inequality, using (20) of Proposition 2, to evaluate the term (analogously to the others: make $\rho = 0$ in $\rho + 2$ or $q_0 = 0$ in the next computations):

$$\begin{aligned}
&2\,t\,const\;(\alpha+1)\,\alpha\,(\alpha-1)\,|\delta| \int_0^t \int_{\mathbf{R}^d} |u|^{\alpha-2+(\rho+1)q_0} |\nabla u|^{\rho+2}\,dxds \\
&\quad= \int_0^t \int_{\mathbf{R}^d} \left[8^{-1}\,p_2\,|u|^{\alpha+1}\right]^{\frac{1}{p_2}} \Big[8^{-1}\,p_3\,D\,(\alpha+1)\,\alpha\,t\,\varepsilon\,|u|^{\mu\varphi+\alpha-1} \\
&\qquad\qquad |\nabla u|^{r+1-\vartheta}\Big]^{\frac{1}{p_3}} \Bigg[\bigg(16^{1-\frac{1}{p_1}}\;((\alpha+1)\,\alpha\,t)^{1-\frac{1}{p_3}}\;const \\
&\qquad\qquad (\alpha-1)\,p_2^{-\frac{1}{p_2}}\,p_3^{-\frac{1}{p_3}}\;D^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\,|\delta|\,\varepsilon^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\bigg)^{p_1} \\
&\qquad\qquad\qquad 2\,D\,\varepsilon\,|u|^{\mu\varphi}|\nabla u|^{r+1-\vartheta}\Bigg]^{\frac{1}{p_1}}\,dxds \\
&\quad\le \frac{1}{8}\int_0^t \int_{\mathbf{R}^d} |u|^{\alpha+1}\,dxds \\
&\qquad+\frac{1}{8}\,D\,(\alpha+1)\,\alpha\,t\,\varepsilon \int_0^t \int_{\mathbf{R}^d} |u|^{\mu\varphi+\alpha-1}|\nabla u|^{r+1-\vartheta}\,dxds \\
&\qquad+\frac{1}{p_1}\bigg(16^{1-\frac{1}{p_1}}\;((\alpha+1)\,\alpha\,t)^{1-\frac{1}{p_3}}\;const\;(\alpha-1)\,p_2^{-\frac{1}{p_2}} \\
&\qquad\qquad p_3^{-\frac{1}{p_3}}\,D^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\,|\delta|\,\varepsilon^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\bigg)^{p_1}\;\|u_0\|^2_{L^2(\mathbf{R}^d)}
\end{aligned}$$

where

$$\begin{aligned}
&\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}=1, \\
&\rho+2=\frac{r+1-\vartheta}{p_1}+\frac{r+1-\vartheta}{p_3}, \\
&\alpha-2+(\rho+1)q_0=\frac{\mu\varphi}{p_1}+\frac{\alpha+1}{p_2}+\frac{\mu\varphi+\alpha-1}{p_3},
\end{aligned}$$

so that

$$\begin{aligned}
&\frac{1}{p_1}+\frac{1}{p_3}=\frac{\rho+2}{r+1-\vartheta}, \qquad \frac{1}{p_2}=1-\frac{\rho+2}{r+1-\vartheta}, \\
&\frac{1}{p_3}=\frac{1}{\alpha-1}\left((\rho+1)q_0-3+(\alpha+1-\mu\varphi)\;\frac{\rho+2}{r+1-\vartheta}\right), \\
&\frac{1}{p_1}=\frac{1}{\alpha-1}\left(3-(\rho+1)q_0+(\mu\varphi-2)\;\frac{\rho+2}{r+1-\vartheta}\right).
\end{aligned}$$

Define

$$H_\alpha\left(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\right) := 4\,t^{\frac{p_1}{p_2}}\,\frac{16^{p_1-1}}{p_1}\,((\alpha+1)\,\alpha)^{1+\frac{p_1}{p_2}}\left(const\,(\alpha-1)\,p_2^{-\frac{1}{p_2}}\right.$$
$$\left. p_3^{-\frac{1}{p_3}}\,D^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\,|\delta|\,\varepsilon^{-\left(\frac{1}{p_1}+\frac{1}{p_3}\right)}\right)^{p_1}\,\|u_0\|^2_{L^2(\mathbf{R}^d)}\,,$$

and remark that if we want to keep $|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}$ bounded as $\varepsilon, |\delta| \searrow 0$, then the $H_\alpha\left(|\delta|\ \varepsilon^{-\frac{2}{r+1-\vartheta}}\right)$, the analogous we obtain making $\rho = 0$ in $\rho+2$, is an infinitesimal.

After re-adding the four terms, we rule out from (22), both,

$$\int_0^t\int_{\mathbf{R}^d}|u|^{\alpha+1}\,dxds + D\,(\alpha+1)\,\alpha\,t\,\varepsilon\int_0^t\int_{\mathbf{R}^d}|u|^{\mu\varphi+\alpha-1}|\nabla u|^{r+1-\vartheta}\,dxds$$
$$\leq 2t\left(2\,\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\left(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\right)\right),$$

$$const\,(\alpha+1)\,\alpha\,(\alpha-1)\,|\delta|\int_0^t\int_{\mathbf{R}^d}|u|^{\alpha-2}\left(1+|u|^{(\rho+1)q_0}\right)|\nabla u|^2\left(1\right.$$
$$\left.+|\nabla u|^\rho\right)dxds \leq \|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\left(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\right).$$

We can finally return over (21) to deduce that

$$\int_{\mathbf{R}^d}|u(t)|^{\alpha+1}\,dx + (\alpha+1)\,\alpha\,\varepsilon\int_0^t\int_{\mathbf{R}^d}|u|^{\alpha-1}\,\nabla u\cdot b(u,\nabla u)\,dxds$$
$$\leq 2\,\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\left(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\right).$$

To conclude, we need explicit the domain of applicability of the Young's inequalities we made. Taking into account that $0 \leq \vartheta < r+1$ and $\rho > 0$ (formally $\rho = 0$ to the analogous)

$$0 \leq \frac{\rho+2}{r+1-\vartheta} \leq 1 \iff r+1-\vartheta \geq \rho+2\,;$$
$$\frac{1}{p_1} \geq 0 \iff \frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{\rho+2}, \quad \text{if } (\rho+1)q_0 \leq 3$$
$$\frac{1}{p_1} \geq 0 \iff \frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{2}, \quad \text{if } (\rho+1)q_0 \geq 3$$
$$\frac{1}{p_3} \geq 0 \iff \alpha+1 \geq \mu\varphi + 3\,\frac{r+1-\vartheta}{2}\,.$$

Therefore we proved at once the uniform L^q bound, no matter how large it can be.

Proposition 3 *Assume that* (A_3), (A_4) *and* (A_5) *or, if* $q_0 = 0$, $(\tilde{A}_5)$ *holds with* $r + 1 - \vartheta \geq \rho + 2$,

$\frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{\rho+2}$, *if* $(\rho+1)q_0 \leq 3$; $\frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{2}$, *if* $(\rho+1)q_0 \geq 3$.

If $u_0 \in L^2(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ *with* $q > 3$ *and* $\varepsilon > 0$, *then the solutions of* (13) *verifies for all* α *such that* $\mu\varphi + 3\,\frac{r+1-\vartheta}{2} \leq \alpha + 1 \leq q$ *and for all* $t \geq 0$

$$\begin{aligned}&\int_{\mathrm{R}^d} |u(t)|^{\alpha+1} dx + (\alpha+1)\,\alpha\,\varepsilon \int_0^t \int_{\mathrm{R}^d} |u|^{\alpha-1}\, \nabla u \cdot b(u, \nabla u)\, dxds \\ &\quad \leq 2\, \|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\Big(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big),\end{aligned} \tag{23}$$

$$\begin{aligned}&\varepsilon \int_0^t \int_{\mathrm{R}^d} |u|^{\mu\varphi+\alpha-1}\, |\nabla u|^{r+1-\vartheta}\ dxds \\ &\quad \leq \frac{1}{D\,(\alpha+1)\,\alpha}\ \Big(2\, \|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\Big(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big)\Big).\end{aligned} \tag{24}$$

$$\begin{aligned}&|\delta| \int_0^t \int_{\mathrm{R}^d} |u|^{\alpha-2}\left(1 + |u|^{(\rho+1)q_0}\right) |\nabla u|^2 \left(1 + |\nabla u|^\rho\right) dxds \\ &\quad \leq \frac{const}{(\alpha+1)\,\alpha\,(\alpha-1)}\ \Big(\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\Big(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big)\Big).\end{aligned} \tag{25}$$

Proposition 4 *Assume that* (A_3), $(\tilde{A}_4)$, $(\tilde{A}_5)$ *holds with* $r + 1 - \vartheta \geq \rho + 2$ *and* $\frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{\rho+2}$. *If* $u_0 \in L^2(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ *with* $q > 3$ *and* $\varepsilon > 0$, *then the solutions of* (13) *verifies for all* α *such that* $\mu\varphi + (3 - (\rho+1)q_0)\,\frac{r+1-\vartheta}{\rho+2} \leq \alpha + 1 \leq q$ *and for all* $t \geq 0$ *the formulae* (23), (24) *and*

$$\begin{aligned}&|\delta| \int_0^t \int_{\mathrm{R}^d} |u|^{\alpha-2+(\rho+1)q_0}\, |\nabla u|^{\rho+2}\, dxds \\ &\quad \leq \frac{const}{(\alpha+1)\,\alpha\,(\alpha-1)}\ \Big(\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\Big(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big)\Big).\end{aligned} \tag{26}$$

Proposition 5 *Assume that* (A_3), $(\tilde{A}_4)$ *and* (A_5) *or, if* $q_0 = 0$, $(\tilde{A}_5)$ *holds with* $r + 1 - \vartheta \geq \rho + 2$, $\frac{\mu\varphi-2}{r+1-\vartheta} \geq \frac{(\rho+1)q_0-3}{\rho+2}$. *If* $u_0 \in L^2(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ *with* $q > 3$ *and* $\varepsilon > 0$, *then the solutions of* (13) *verifies for all* α *such that* $\mu\varphi + 3\,\frac{r+1-\vartheta}{\rho+2} \leq \alpha + 1 \leq q$ *and for all* $t \geq 0$ *the formulae* (23), (24) *and*

$$\begin{aligned}&|\delta| \int_0^t \int_{\mathrm{R}^d} |u|^{\alpha-2}\ \left(1 + |u|^{(\rho+1)q_0}\right)\ |\nabla u|^{\rho+2}\, dxds \\ &\quad \leq \frac{const}{(\alpha+1)\,\alpha\,(\alpha-1)}\ \Big(\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathbf{R}^d)} + H_\alpha\Big(|\delta|\ \varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big)\Big).\end{aligned} \tag{27}$$

Proposition 6 *Assume that* (A_3), (A_4), $(\tilde{A}_5)$ *holds with* $r + 1 - \vartheta \geq \rho + 2$. *If* $u_0 \in L^2(\mathbf{R}^d) \cap L^q(\mathbf{R}^d)$ *with* $q > 3$ *and* $\varepsilon > 0$, *then the solutions of* (13) *verifies for all* α *such that*

if $(\rho+1)q_0 \le 3$, $\frac{\mu\varphi-2}{r+1-\vartheta} \ge \frac{(\rho+1)q_0-3}{\rho+2}$ *and* $\mu\varphi + (3-q_0)\,\frac{r+1-\vartheta}{2} \le \alpha+1 \le q$*;*
if $q_0 < 3 < (\rho+1)q_0$, $\frac{\mu\varphi-2}{r+1-\vartheta} \ge \frac{(\rho+1)q_0-3}{2}$ *and* $\mu\varphi + (3-q_0)\,\frac{r+1-\vartheta}{2} \le \alpha + 1 \le q$*;*
if $q_0 \ge 3$, $\frac{\mu\varphi-2}{r+1-\vartheta} \ge \frac{(\rho+1)q_0-3}{2}$ *and* $\mu\varphi + (3-q_0)\,\frac{r+1-\vartheta}{\rho+2} \le \alpha+1 \le q$*;*

and for all $t \ge 0$ *the formulae* (23), (24) *and*

$$|\delta| \int_0^t \int_{\mathrm{R}^d} |u|^{\alpha-2+q_0}\,\left(1+|u|^{\rho q_0}\right)\,|\nabla u|^2\,\left(1+|\nabla u|^{\rho}\right)\,dxds \le \frac{const}{(\alpha+1)\,\alpha\,(\alpha-1)}\left(\|u_0\|^{\alpha+1}_{L^{\alpha+1}(\mathrm{R}^d)} + H_\alpha\Big(|\delta|\,\varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\Big)\right). \tag{28}$$

3 Convergence Proof

We prove here Theorem 2 (we proceed the same way to prove the three others).

Proof We begin proving (7). By (23) of Proposition 4 we can use the Young measure representation theorem in L^q (Lemma 1) and its limit representation formula (10) to show that ν satisfies (7). Still, using a standard regularization of $\operatorname{sgn}(u-k)(f(u)-f(k))$ and $|u-k|$ ($k \in \mathbf{R}$), which satisfies the growth condition (9) in the representation theorem, we see it is sufficient to show that there exists a bounded measure $\mu \le 0$ such that

$$\partial_t \eta(u) + \operatorname{div} q(u) \longrightarrow \mu \quad \text{in} \quad \mathscr{D}'(\mathbf{R}^d \times (0,T))$$

for an arbitrary convex function η (we assume η', η'', η''' to be bounded functions on $\mathbf{R}$). To do this reconsider formulae (14)–(15):

$$\partial_t \eta(u) + \operatorname{div} q(u) = \mu_1 + \mu_2 + \mu_3$$

with the notation

$$\begin{aligned}
\mu_1 &:= \varepsilon\ \operatorname{div}\big(\eta'(u)\ b(u,\nabla u)\big)\,,\\
\mu_2 &:= -\varepsilon\,\eta''(u)\ \nabla u \cdot b(u,\nabla u),\\
\mu_3 &:= \delta\,\eta'''(u)\nabla u \cdot \Big(\big[c_{jk}\,(\nabla G(u))\,\big]\nabla G(u) - C\,(\nabla G(u))\Big)\\
&\quad + \delta \sum_{j,k} \partial^2_{x_j x_k}\big(\,\eta'(u)\,g(u)\,c_{jk}\,(\nabla G(u))\big)\\
&\quad + \delta\ \operatorname{div}\,\Big(\eta''(u)\big(\,C\,(\nabla G(u)) - \big[c_{jk}\,(\nabla G(u)) + c^T_{jk}\,(\nabla G(u))\big]\nabla G(u)\big)\Big)\\
&\quad - \delta\ \operatorname{div}\,\Big(\,\eta'(u)\big[c_{jk}\,(\nabla G(u))\,\big]\nabla g(u)\Big)\,,
\end{aligned}$$

without the last term in the case where $q_0 = 0$.

For each positive $\theta \in \mathrm{C}^\infty_0(\mathbf{R}^d \times (0,T))$ we evaluate $\langle \mu_i, \theta\rangle$ for $i = 1, 2, 3$.

By assumption (A_2) we have, with K suitable constants,

$$\begin{aligned}|\langle \mu_1, \theta\rangle| &\le \varepsilon \int_0^T \int_{\mathbf{R}^d} \left|\nabla\theta \cdot \eta'(u)\, b(u, \nabla u)\right| \, dxdt \\ &\le K\,\varepsilon \int_0^T \int_{\mathbf{R}^d} |\nabla\theta| \, dxdt \; + K\,\varepsilon \int_0^T \int_{\mathbf{R}^d} |\nabla\theta|\, |u|^{\mu}\, |\nabla u|^{r}\, dxdt\end{aligned}$$

and, if we use Hölder's inequality within (20) of Proposition 2 or (24) of Proposition 4, we get

$$\begin{aligned}|\langle \mu_1, \theta\rangle| \le\;& K\,\varepsilon\, \|\nabla\theta\|_{L^1(\mathbf{R}^d\times(0,T))} \\ &+ K\,\varepsilon^{\frac{1-\vartheta}{r+1-\vartheta}} \left[\varepsilon \int_0^T \int_{\mathbf{R}^d} |u|^{\mu\varphi+\alpha-1}\; |\nabla u|^{r+1-\vartheta}\; dxdt\right]^{\frac{r}{r+1-\vartheta}} \\ &\qquad\left[\int_0^T \int_{\mathbf{R}^d} |\nabla\theta|^{\frac{r+1-\vartheta}{1-\vartheta}}\; |u|^{z}\; dxdt\right]^{\frac{1-\vartheta}{r+1-\vartheta}} \\ \le\;& K\,\varepsilon\, \|\nabla\theta\|_{L^1(\mathbf{R}^d\times(0,T))} \\ &+ K\,\varepsilon^{\frac{1-\vartheta}{r+1-\vartheta}} \left[\int_0^T \int_{\mathbf{R}^d} |\nabla\theta|^{\frac{r+1-\vartheta}{1-\vartheta}}\; |u|^{z}\; dxdt\right]^{\frac{1-\vartheta}{r+1-\vartheta}}\end{aligned}$$

where $z = \mu\left(1 + r\,\frac{1-\varphi}{1-\vartheta}\right) - r\,\frac{\alpha-1}{1-\vartheta}$, which we can keep equal to zero as $\alpha - 1$ varies along $[0, q-2]$ and when $\alpha - 1$ reaches the maximum $q - 2$, then z must run over $[0, q]$, i.e., provided that $(\mu - q)\,(r + 1 - \vartheta) \le r\,(\mu\varphi - 2)$, we use (18) of Proposition 2 or (23) of Proposition 4 to conclude with appropriate L^p that

$$|\langle \mu_1, \theta\rangle| \le K\,\varepsilon\, \|\nabla\theta\|_{L^1(\mathbf{R}^d\times(0,T))} + K\,\varepsilon^{\frac{1-\vartheta}{r+1-\vartheta}}\; \|\nabla\theta\|_{L^p(\mathbf{R}^d\times(0,T))}.$$

About μ_2, η is convex and $\nabla u \cdot b(u, \nabla u) \ge 0$ because of (A_3), thus

$$\langle \mu_2, \theta\rangle = -\,\varepsilon \int_0^T \int_{\mathbf{R}^d} \theta\;\eta''(u)\;\nabla u \cdot b(u, \nabla u)\; dxdt \le 0$$

and by (18) of Proposition 2 it is bounded.

For μ_3, we have by assumptions (A_4) and (A_5) that

$$\begin{aligned}|\langle \mu_3, \theta\rangle| \le K\,|\delta| &\int_0^T \int_{\mathbf{R}^d} \left|\eta'''(u)\right| \theta\, |u|^{(\rho+1)q_0}\, |\nabla u|^{\rho+2}\; dxdt \\ &+ K\,|\delta| \int_0^T \int_{\mathbf{R}^d} \left|\eta'(u)\right| |D^2\theta|\; |u|^{(\rho+1)q_0}\, |\nabla u|^{\rho}\; dxdt \\ &+ K\,|\delta| \int_0^T \int_{\mathbf{R}^d} \left|\eta''(u)\right| |\nabla\theta|\, |u|^{(\rho+1)q_0}\, |\nabla u|^{\rho+1}\; dxdt\end{aligned}$$

$$
\begin{aligned}
&+ K\,|\delta| \int_0^T \int_{\mathbf{R}^d} \left|\eta'(u)\right| |\nabla\theta|\, |u|^{(\rho+1)q_0-1}\, |\nabla u|^{\rho+1}\ dxdt \\
&+ K\,|\delta| \int_0^T \int_{\mathbf{R}^d} \left|\eta'(u)\right| |\nabla\theta|\, |u|^{\rho q_0}\, |\nabla u|^{\rho+1}\ dxdt,
\end{aligned}
$$

without the last two terms in the case where $q_0 = 0$ or without the last one if we assume (A_5). We use Hölder's inequality within Proposition 4 as follows:

$$
\begin{aligned}
|\langle \mu_3, \theta\rangle| \leq\ & K\,|\delta|\,\varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}\ \|\theta\|_{L^{1-\frac{\rho+2}{r+1-\vartheta}}(\mathbf{R}^d\times(0,T))} \\
& \left[\varepsilon \iint |u|^{\mu\varphi+\alpha_1-1}\,|\nabla u|^{r+1-\vartheta}\ dxdt\right]^{\frac{\rho+2}{r+1-\vartheta}} \\
& + K\,|\delta|\,\varepsilon^{-\frac{\rho}{r+1-\vartheta}}\ \|D^2\theta\|_{L^{1-\frac{\rho}{r+1-\vartheta}}(\mathbf{R}^d\times(0,T))} \\
& \left[\varepsilon \iint |u|^{\mu\varphi+\alpha_2-1}\,|\nabla u|^{r+1-\vartheta}\ dxdt\right]^{\frac{\rho}{r+1-\vartheta}} \\
& + K\,|\delta|\,\varepsilon^{-\frac{\rho+1}{r+1-\vartheta}}\ \|\nabla\theta\|_{L^{1-\frac{\rho+1}{r+1-\vartheta}}(\mathbf{R}^d\times(0,T))} \\
& \left[\varepsilon \iint |u|^{\mu\varphi+\alpha_3-1}\,|\nabla u|^{r+1-\vartheta}\ dxdt\right]^{\frac{\rho+1}{r+1-\vartheta}} \\
& + K\,|\delta|\,\varepsilon^{-\frac{\rho+1}{r+1-\vartheta}}\ \|\nabla\theta\|_{L^{1-\frac{\rho+1}{r+1-\vartheta}}(\mathbf{R}^d\times(0,T))} \\
& \left[\varepsilon \iint |u|^{\mu\varphi+\alpha_4-1}\,|\nabla u|^{r+1-\vartheta}\ dxdt\right]^{\frac{\rho+1}{r+1-\vartheta}} \\
& + K\,|\delta|\,\varepsilon^{-\frac{\rho+1}{r+1-\vartheta}}\ \|\nabla\theta\|_{L^{1-\frac{\rho+1}{r+1-\vartheta}}(\mathbf{R}^d\times(0,T))} \\
& \left[\varepsilon \iint |u|^{\mu\varphi+\alpha_5-1}\,|\nabla u|^{r+1-\vartheta}\ dxdt\right]^{\frac{\rho+1}{r+1-\vartheta}}
\end{aligned}
$$

where

$$\alpha_1 + 1 = \frac{r+1-\vartheta}{\rho+2}\,(\rho+1)q_0 + 2 - \mu\varphi,$$

$$\alpha_2 + 1 = \frac{r+1-\vartheta}{\rho}\,(\rho+1)q_0 + 2 - \mu\varphi,$$

$$\alpha_3 + 1 = \frac{r+1-\vartheta}{\rho+1}\,(\rho+1)q_0 + 2 - \mu\varphi,$$

$$\alpha_4 + 1 = \frac{r+1-\vartheta}{\rho+1}\,((\rho+1)q_0 - 1) + 2 - \mu\varphi,$$

which worst restriction is given by

$$\frac{\mu\varphi}{r+1-\vartheta} \leq \frac{(\rho+1)q_0}{\rho+2}$$

and

$$\alpha_5 + 1 = \frac{r+1-\vartheta}{\rho+1}\,\rho q_0 + 2 - \mu\varphi,$$

subordinated to the more severe condition of

$$\frac{\mu\varphi}{r+1-\vartheta} \leq \frac{\rho q_0}{\rho+1}.$$

Therefore,

$$|\langle \mu_3, \theta\rangle| \leq \; K\,|\delta|\,\varepsilon^{-\frac{\rho+2}{r+1-\vartheta}}.$$

Finally the condition $\delta = o\left(\varepsilon^{\frac{\rho+2}{r+1-\vartheta}}\right)$ is sufficient to the conclusion. To show (8) we can follow the same argument as in Correia-LeFloch [1].

Acknowledgements The author was partially supported by the Grant Agency FCT of Portugal under program PEst-OE/MAT/UI0117/2014.

References

1. Correia, J.M.C., LeFloch, P.G.: Nonlinear diffusive-dispersive limits for multidimensional conservation laws. In: Advances in Partial Differential Equations and Related Areas Beijing (1997), pp. 103–123. World Sci. Publ., River Edge, NJ, (1998). arXiv:0810.1880
2. Kružkov, S.N.: First order quasilinear equations in several independent variables. Mat. Sb. **81**, 285–355 (1970); Math. USSR Sb. **10**, 217–243 (1970)
3. Lax, P.D., Levermore, C.D.: The small dispersion limit of the Korteweg-de Vries equation. Comm. Pure Appl. Math. **36**, I, p. 253, II, p. 571, III, p. 809 (1983)
4. LeFloch, P.G., Natalini, R.: Conservation laws with vanishing nonlinear diffusion and dispersion. Nonlinear Anal. **36**, 213–230 (1999)
5. Perthame, B., Ryzhik, L.: Moderate dispersion in conservation laws with convex fluxes. Comm. Math. Sci. **5**(2), 473–484 (2007)
6. Schonbek, M.E.: Convergence of solutions to nonlinear dispersive equations. Comm. Part. Diff. Equa. **7**, 959–1000 (1982)
7. Slemrod, M.: From Boltzmann to Euler: Hilbert's 6th problem revisited. Comput. Math. Appl. **65**, 1497–1501 (2013)
8. Whitham, G.B.: Linear and Nonlinear Waves. Pure anD Applied Mathematics. Wiley-Interscience Publication, New York (1974)

Modelling Consumer Preferences for Novel Foods: Random Utility and Reference Point Effects Approaches

Irina Dolgopolova, Ramona Teuber, Viola Bruschi, Gerhard-Wilhelm Weber, Nina Danilenko and Efim Galitskiy

Abstract Advances in the bioeconomy lead to a range of innovative products appearing at the consumer markets. However, these products often face consumer resistance. In this chapter we test if a reference point effects approach can provide more information about consumers decision-making regarding novel food products than a random utility approach. We draw on data from a survey and second-price Vickrey auction for novel foods with health and environmental benefits. First, we analyze consumer choices within a random utility framework and compare stated and revealed preferences. Second, reference point effects are included into the methodological framework and weighted and unweighted models for revealed preferences are obtained. Results of the random utility estimations provide information on attributes value and the evidence of overestimated stated preferences. The reference point approach indicates the presence of reference points in the experimental auction data and asymmetrical effects of gains and losses on utility values.

I. Dolgopolova (✉)
Technical University of Munich, Munich, Germany
e-mail: irina.dolgopolova@tum.de

R. Teuber
University of Copenhagen, Copenhagen, Denmark
e-mail: rt@ifro.ku.dk

V. Bruschi
Leibniz Institute of Agricultural Development in Transition Economies, Halle, Germany
e-mail: bruschi@iamo.de

G.-W. Weber
Middle East Technical University, Ankara, Turkey
e-mail: gweber@metu.edu.tr

N. Danilenko
Baikal State University, Irkutsk, Russia
e-mail: nina.danilenko@gmail.com

E. Galitskiy
Higher School of Economics, Moscow, Russia
e-mail: egalit@yandex.ru

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_10

Keywords Reference point · Novel food · Stated · Revealed preferences

1 Introduction

When environmental changes and bio-technological advances lead to the transformation of end-products deeper understanding of consumers decision-making is required to obtain more knowledge on resistance to innovations. Advances in the bioeconomy result in the development of innovative products ranging from biofuels to vaccines or new packaging material. In agriculture and, more specifically in crop-breeding, biotechnologies lead to the development of new foods that might not only be beneficial for the environment but also for consumers health and well-being.

However, the introduction of novel products into consumer market often faces consumer resistance. Taking novel foods as an example, it has been observed that consumers acceptance of foods with health benefits depends on a variety of factors [8, 33] and that a certain health benefit is not necessarily valued strongly by consumers [38]. Among possible reasons for novel food product market failures are: low trustworthiness or knowledge about foods with health benefits, unwillingness to pay higher prices, and concerns about taste and naturalness [26]. Moreover, technologies not directly related to the production of a certain product that are negatively perceived by some consumer groups (like genetic modification (GM)) can be a source of resistance to innovations for all the products that have any kind of technological transformation (not even necessarily GM) involved [8].

Changes at the consumer markets that follow the advances in bioeconomy require knowledge about consumer preferences towards novel foods and about the reasons for novelty resistance. From a methodological point of view, it requires deviation from traditional models of consumer behavior that sometimes fail to explain consumers decision-making towards more flexible models that incorporate previously unaccounted for factors.

The analysis of consumer preferences is traditionally based on the assumption of rationality. However, research on departures from rationality in decision-making [13, 34, 35] allows for modelling of psychological effects that can provide new insights in consumer decision making concerning the acceptance of novel products. This study combines recent advances in the applications of the Kahneman and Tversky [35] reference dependence approach with traditional modelling of consumer choices. Our work will draw upon one of the possible cognitive anomalies in decision-making process, i.e. reference points. Applied to food choice decisions, the reference point effects approach can provide evidence on previously not accounted for determinants of consumer choices.

Experimental evidence supporting distortions of rationality in consumer behavior have been mainly obtained from specifically designed procedures [3, 5]. Additionally, evidence supporting the existence of reference points in consumers choices was found in real market data [10, 12, 15]. However, no attempts have been made so far to include the implications of loss aversion and reference point effects into the analysis of experimental auctions data on consumer choices of novel foods.

Novel functional foods belong to the category of credence goods. Consumer's perception of credence goods is complicated by the inability to measure quality levels before or after the purchase or even consumption. In the context of novel foods with health benefit when consumers are dealing with high levels of uncertainty, the existence of reference points might provide more information for explaining the heterogeneity among consumer choices. As pointed out by Hu, Adamowicz, and Veeman [12], heterogeneity in consumer attitudes may come not only from traditionally considered characteristics like socio-demographics, knowledge, taste, and product attributes but also from the framing of the decision process. From the reference dependence perspective heterogeneity might also arise from the influence of reference prices as well as the influence of gains and losses.

In a broader microeconomic context, it also raises questions related to consumers preferences formation. Consumers valuations might not only arise directly from the possible health effects. It is also possible to assume that consumers might value health benefit in the context of the prices they usually pay for the products of the same category or their preferences for such products.

Thus, this chapter contributes to the literature on modeling consumers preferences by comparing a random utility approach with a reference point effects approach. We employ data from experimental auctions designed to measure willingness-to-pay (WTP) for novel bakery foods with health and environmental benefits. Our aim is to understand if the application of reference point effects will provide new insights on the process of consumers preferences for such foods.

Our chapter is organized as follows. Section 2 provides an overview of the literature on reference point effects in consumer choices. Section 3 describes our experimental setting and the results of the bidding procedure. Section 4 consists of two parts: the first part describes the results of the traditional random utility approach to the data, while the second part presents results from the reference point effects approach. Conclusions are presented in Sect. 5.

2 Literature Survey on Reference Dependence Approach to Modeling Consumer Behavior

The reference point effects approach to consumer choices belongs to a relatively new area of economic and marketing research that takes root in prospect theory by Kahneman and Tversky [13]. The basic elements of this approach include the existence of reference points, gains and losses, and the effect of loss aversion. Reference points can be represented by price, quality or activity that choice alternatives are compared to. Gains and losses are, respectively, positive and negative departures from a reference point. Loss aversion implies that losses outweigh gains in choice decisions. The range of possible roles that reference prices can play in consumer decision-making is discussed at length in Cheng and Monroe [7] and Lee [16].

Reference point effects have been incorporated into models of consumer choice by several scholars. A theoretical and empirical approach to incorporate reference point effects into traditional economic theory of consumer choice has been performed by Putler [27]. He specifies a utility function that includes gains and losses and discusses implications for traditional economic and marketing paradigms of consumer choice. When applied to retail data on egg sales, Putler's theoretical framework provides significant results for reference price effects and asymmetric responses on gains and losses.

Lattin and Bucklin [15] investigated reference effects of price and promotion. They formalize promotional reference point as consumers prior exposure to the promotional activities of a specific brand and price reference point as consumers exposure to the price of a specific brand on previous purchases. Applied to scanner panel data on ground coffee, the model of consumer response proves significant reference effects of promotional activity.

Kalwani and Yim [14] suggest operationalizing price gains and losses using the expected prices directly elicited from consumers. Dummy variable representing gains takes the value of 1 in case if expected price exceeds retail price and 0 otherwise. The loss variable is constructed as being equal to 1 if a retail price exceeds expected price and 0 otherwise. They observe the outweighing effects of losses compared to the effects of gains.

Modelling reference points for both price and quality of orange juice based on scanner panel data has been done by Hardie, Johnson, and Fader [10]. The authors operationalize gains as the amount by which quality or price of a specific brand exceeds that of the reference brand, and losses as the amount by which quality or price of a specific brand is below that of a reference brand. Incorporating reference points into a multinomial logit model proves consistent with loss aversion and decreases heterogeneity and nonstationarity of the model.

Reference points have also been included into models of food attribute demand. Hu, Adamowicz, and Veeman [12] analyze reference points for price and genetically modified ingredients in pre-packaged sliced bread. Reference points are obtained from questions on consumer's regular bread purchases conducted before the discrete choice experiment. They observe strong reference point effects, especially for the price. The economic implications of reference point effects include possible changes in welfare measures of consumer choices.

More recently, Hess, Rose, and Hensher [11] found support for the prospect theory view of decision making when applied to car travel data from discrete choice experiments. They offer evidence of framing effects in respondents decision-making, so that preferences are formed not relative to the absolute values of the attributes, but relative to differences in values according to a specific reference point.

However, testing for the presence of reference point effects has not always provided significant results. The universality of loss aversion in consumer choices has been questioned in the scanner panel data analysis by Bell and Lattin [6]. They employed data on refrigerated orange juice and additional 11 product categories to test the reference dependent model. They found no asymmetric price response when applying a brand specific reference dependent model model. They also reported decreasing evidence of loss version in the reference dependent model.

In general, most studies indicate that accounting for reference effects in consumer choice provides additional evidence on the framing of consumer decision-making.

3 Experimental Design and Results

In this chapter we use data from a survey followed by experimental auctions (second price sealed bid Vickrey auctions [39]), designed to determine consumers willingness-to-pay for healthy and environmental attributes in novel food that contributes to preserving biodiversity.

The cereal food products that were presented to the consumers are not yet marketed novel foods with health benefits: a bread roll and 130 g pack of biscuits. Both products are produced from an old wheat variety that contains Anthocyanin, a natural substance of purplish color that is potentially beneficial for health. Old wheat varieties also help in preserving biodiversity. Due to the novelty in content and appearance, consumer preferences for Anthocyanin-rich cereal products are formed during the auction as respondents get information about the properties and potential benefits of the products. During the auction, participants submitted bids for the products with health and environmental attributes. Potential market prices of the products were not revealed to the consumers during the auctions.

The possibility to elicit non-hypothetical monetary values of healthy attributes is the main advantage of an experimental auction procedure [18, 31]. The second-price Vickrey auction format is an incentive compatible and willingness-to-pay revealing mechanism [25], which is also effective in measuring consumers willingness to pay for quality differences in food products [37]. However, compared to other auction mechanisms it might generate higher valuations [18].

Experiments were organized at the campuses of two universities in Russia: at the Higher School of Economics in Moscow and at the Baikal State University and Law in Irkutsk in December 2013. Participants included mainly students (see Table 1). Due to the fact that in this chapter we are mainly interested in methodological issues the sampling composition does not place a burden on the analysis.

The procedure consisted of the following stages. At first, participants were asked to complete a survey in which they provided demographic information and answered questions about their preferences and concerns for food products to be presented during the auction, beliefs about connection between food and health, knowledge about anthocyanin and old wheat varieties and their willingness-to-pay for products with health benefits.

The auction itself was divided into few steps. First, participants were provided with 200 Rubles (approx. 5 Euros) incentive and were asked to bid for a candy bar to familiarize with the auction procedure. Second, after familiarizing participants with the procedure they were asked to post bids for anthocyanin-rich bread roll

Table 1 Characteristics of participants (N=212)

Variable	Definition	Mean	St.dev.
Gender	1-male; 2-female	1.726	0.447
Age	Age in years	22.322	6.617
Education	Educational level 1-BS; 2-MS; 3-PhD	1.224	0.451
Income (Russian rubles)	1-less than 30000; 2- from 30001 to 60000; 3- from 60000 to 90000; 4-more than 90000	2.633	1.006
Nutrition-related illnesses	1 = yes; 2 = no	1.726	0.447
Sport activities	1 = yes; 2 = no	1.604	0.490
Smoking	1 = yes; 2 = no	1.811	0.392
Alcohol consumption	1-every day; 2-few times a week; 3-few times a month; 4-few times a year; 5-never	3.526	0.818

and biscuits. We performed three bidding rounds for each product. Considering the already significant length of the procedure, we did not include repetitions to avoid participant's fatigue as there exist evidence that repetition might not have an improving effect on the bids [2].

In the first round only basic information indicating that the product presented for visual inspection does not have any particular characteristics were provided. Participants submitted their sealed bids for the base product. Then, participants were informed about one of the two attributes (health or environmental) and asked to submit bids again. In the last round, one more attribute was added and participants were asked again to post the bids. After that the whole procedure was repeated with another product. The order of the attributes introduced was randomized. At the end, one auction round was randomly chosen as binding, the winner of this round was identified and the audience was informed about the winning price.

We chose a sequential order for the attribute introduction as it uncovers the evolution of the bids in reaction to the attributes introduced. Previously, a sequential order for eliciting willingness-to-pay was performed by Marette et al. [22] to determine the effect of risk and benefit information on the choice of fish species, by Marette et al. [21] for evaluating the effect of health information on consumers choices of functional food, and by Rozan [28] to estimate WTP for food safety.

The results of the experiment produced four main data points: (i) hypothetical price premiums for healthy foods from the questionnaire, (ii) bids at first no information round; (iii) bids at the second + one attribute round, and (iv) bids in the third + two attributes round. Combining survey and auction data allows us to compare stated and revealed preferences following Adamowicz, Louviere, and Williams [1].

A basic analysis of experimental data suggests that bids for both products increased from round to round. Descriptive statistics for the bids is presented in Table 2. The levels of mean bid increases for additional attributes are quite substantial. This evidence is supported by a Wilcoxon test of pairs of means ($p < 0.01$). We can assume that preferences for each additional attribute are distributed so that

Table 2 Descriptive statistics of bids (in Russian Rubles)

	Bread			Biscuits		
	Base	One attr.	Two attr.	Base	One attr.	Two attr.
Mean	14.24	25.54	32.75	21.19	27.03	38.02
St.dev.	21.23	27.99	34.55	17.31	22.67	41.23
Number of zero bids	58	25	22	30	29	25

a product with two attributes is preferred to a product with only one attribute. The number of zero bids is decreasing in each round as new information about health attributes is received. It indicates that additional attributes can initiate a purchase for some consumers who were not interested in the grain products with basic characteristics.High values of standard deviations can be attributed to high level of heterogeneity between consumers.

4 Analysis of Consumer Choice

4.1 Random Utility Approach

The traditional way to obtain welfare measures on consumer choices is the random utility approach [9], in which utility from a choice alternative consists of deterministic and stochastic parts:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \tag{1}$$

where U_{ij} is the utility level of the product j for consumer i; V_{ij} is the deterministic part of consumer is utility, and ε_{ij} is the stochastic part. During experimental auction consumer solves the choice problem by submitting a bid which value depends on the level of utility that can be obtained from a product. Higher utility levels correspond to the intention to pay price premium. Each possible choice available to the consumer can be described as the linear random utility model, in our notation:

$$V_{ij} = \alpha_j p + \beta_j z_i + \gamma_j w_{ij} + \varepsilon_{ij}, \tag{2}$$

where p is the market price of the product; z_i stands for individual i's characteristics; w_{ij} represents the attributes of the alternative product; ε_{ij} denotes unobservable stochastic Gumbel distributed random term. Then, for each individual i, the utilities of possible choices can be described as: V^0_{ij} for the base product and V^1_{ij} for the product with a health benefit. The respondents choice alternative V^1_{ij} is denoted $Y = 1$ and states that $U^1_{ij} > U^0_{ij}$. If a respondent chooses V^0_{ij}, then $Y = 0$

and $U_{ij}^1 \leq U_{ij}^0$. Then,the probability of choosing the option with higher utilitycan be described as:$Prob(Y = 1 \mid \mathbf{z}_i, \mathbf{w}_j^0, \mathbf{w}_j^1) = Prob(\mathbf{x}'\beta + \varepsilon > 0 \mid \mathbf{x})$, where all the observable differences between two utility functions are collected in $\mathbf{x}'$, β, and ε_{ij} summarizes the differences in random elements. The logistic distribution model is: $Prob(Y = 1 \mid \mathbf{x}) = \exp(\mathbf{x}'\beta)/(1 + \exp(\mathbf{x}'\beta))$.

As stated earlier, four major information points about consumer preferences from the experimental auction are available. We also use data from the survey performed before the auction to derive stated preferences for healthy foods. Consequently, we estimate one stated preference and three revealed preference models. A description of the variables included in all estimations is presented in Table 3.

In the stated preference model, the dependent variable is constructed from the yes/no answers to the question: Are you willing to pay price premium for food that can improve your health? Dependent variables for the revealed preferences models are constructed considering the price level of the region. Since Vickrey auctions are demand revealing and participants are assumed to submit bids that reflect their true valuation of the product, we constructed a dependent binary variable which equals 1 if a participant submitted a bid higher than the market price for the base product (i.e., bread roll of biscuits without any additional characteristics), and 0 in any other case. Discretization of the auction results is necessary for the comparability between stated and revealed preferences models. The variable indicates that the consumer maximizes utility if her valuation of the auctioned bread or biscuits is high enough to buy this bread or a pack of biscuits at the market.

Multinomial logit is usually employed for the estimation of additive random utility models, however, in the case of a binary dependent variable the results of the estimation are the same as for the binary logit. The results of the logit regressions including participants characteristics and variables concerning consumer preferences for bread and healthy foods are presented in Table 4. We can see different preference structures underlying stated and revealed choices. Among individual characteristics positively influencing stated preferences for healthy foods are previous purchases of such foods and a perceived connection between food consumption and health. Regarding product-related characteristics the average market price was not statistically significant in the stated preference estimation indicating that respondents considerations of hypothetical price premiums have lower predictive power than the actual revealed preferences. In the case of bids submitted by the participants in the first round (base product) only previous purchases of healthy food in the case of bread and price for both bread and biscuits were significant. For revealed preferences estimations, statistically significant factors of bid increases were the product attributes introduced during the auction and the market price.

Since the coefficients of the logit regressions are not directly interpretable, we obtain contrasts of predictive margins that demonstrate changes in probabilities associated with each level of significant predictors holding other covariates at observed values (see Table 5).

All the significant predictors are dummy variables with only two levels. For the stated preferences estimation previous purchases of healthy foods increase the prob-

Table 3 Description of explanatory variables

Variable	Description
Frequency	How often do you consume bread/biscuits? (5-every day; 4-few times a week; 3-few times a month; 2-few times a year; 1-never)
Price	Please, indicate how important is price for your choice of bread/biscuits (1-most important; 0-otherwise)
Taste	Please, indicate how important is taste for your choice of bread/biscuits (1-most important; 0-otherwise)
Healthiness	Please, indicate how important is health for your choice of bread/biscuits (1-most important; 0-otherwise)
Novelty	Please, indicate how important is novelty for your choice of bread/biscuits (1-most important; 0-otherwise)
Tradition	Please, indicate how important is tradition for your choice of bread/biscuits (1-most important; 0-otherwise)
Connection between food and health	Do you agree that consumption of certain foods can influence your health? (1-yes; 0-otherwise)
Previous purchases of healthy food	Have you ever bought food because it can improve your health status? (1-yes; 0-otherwise)
Consent to pay price premiums for products with healthy attributes	Are you willing to pay price premium for food that can improve your health? (1-yes; 0-otherwise)
Gender	1-male; 0-female
Income (rubles per month)	1-less than 30000; 2-30001-60000; 3-60000-90000; 4-more than 90000
Nutrition-related illnesses	Do you have nutrition-related illnesses? (1-yes; 0-no)
Sport	Do you do sport regularly? (1-yes; 0-no)
Smoke	Do you smoke? (1-yes; 0-no)
Alcohol	How often do you consume alcohol?(5-every day; 4-few times a week; 3-few times a month; 2-few times a year; 1-never)
Attribute value	Dummy variable: 1 if bid in the corresponding round exceeds bid in the previous round; 0 otherwise
Price	Average market price of a corresponding product

Table 4 Coefficients from binary logit estimations (1% significance level)

Utility function variables	Stated preferences		1st round		2nd round		3rd round	
	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits
Factors related to product consumption								
Frequency	−0.25	0.11	0.02	0.25	0.00	0.20	0.10	0.14
Price	−1.57	—	0.53	1.47	−1.27	0.93	−0.17	0.84
Taste	−1.15	−0.10	−0.74	−0.29	−1.04	−0.38	−0.29	0.13
Healthiness	0.35	−0.30	−0.51	−0.33	−0.46	−0.34	−0.19	−0.97
Novelty	−0.42	−0.36	0.51	−0.53	−0.46	−0.48	0.14	−0.11
Tradition	−1.40	−0.68	0.29	—	0.29	0.83	−0.07	−0.07
Connection between food and health	2.31***	2.20***	−1.01	−0.55	−0.36	−1.00	−0.55	−0.17
Previous purchases of healthy food	1.80***	1.89***	1.62***	0.66	0.76	0.88	0.98	1.93***
Consent to pay price premiums for products with healthy attributes	—	—	0.56	0.34	0.46	0.15	0.65	0.17
Socio-demographic and lifestyle characteristics								
Gender	0.70	0.23	−0.18	0.10	−0.23	−0.69	0.17	0.22
Income	0.04	0.05	−0.40	0.07	−0.26	−0.04	−0.20	−0.13
Nutrition-related illnesses	−0.18	−0.48	0.48	−0.76	0.09	−0.71	0.31	0.34
Sport	−0.81	−1.19	0.41	0.38	0.18	−0.10	−0.09	−0.11
Smoke	0.01	0.05	1.18	0.40	0.68	0.42	0.61	0.51
Alcohol	−0.06	0.07	0.27	0.17	0.27	0.23	−0.13	0.58
Product-related characteristics								
Attribute value	—	—	—	—	2.47***	2.24***	1.82***	2.30***
Price	−0.05	−0.06	−0.42***	−0.28***	−0.26***	−0.27***	−0.19***	−0.32***
Constant	1.21	−0.52	3.42	3.41	1.72	3.56	2.12	1.48
AIC	183.752	184.003	215.728	230.861	227.731	210.611	228.801	188.343
BIC	236.364	232.550	271.629	283.310	286.919	269.800	287.990	247.513
Pseudo R^2	0.27	0.24	0.31	0.27	0.28	0.35	0.21	0.40

*, **, and *** denote significance at the .1, .05, .01 level, respectively. Standard errors are reported in parentheses

ability of paying price premiums for healthy attributes in foods by 28 and 33% for bread and biscuits, respectively. In case respondents see a strong link between food consumption and health, the probability of paying price premiums for healthy attributes increases by 30% for both products. For the first round of the auction we observe that for respondents who agree that consumption of certain foods influences

Table 5 Contrasts of predictive margins for statistically significant explanatory variables (1% significance level, covariates at observed values)

	Stated preferences		1st round		2nd round		3rd round	
Utility function variables	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits
Attribute value	—	—	—	—	0.45***	0.37***	0.32***	0.32***
Previous purchases of healthy food	0.28***	0.33***	0.23***	—	—	—	—	0.27***
Connection between food and health	0.38***	0.38***	—	—	—	—	—	—

*, **, and *** denote significance at the .1, .05, .01 level, respectively. Standard errors are reported in parentheses

health the probability of paying the price higher than market price for the bread base product is 23% higher. For the second and the third rounds of the auction, the highest increase in the probability of paying higher prices is provided by the introduction of additional product attributes. Previous purchases of healthy foods were a significant predictor for the purchase of biscuits in the third auction round with an increase in probability of 27%. The introduction of positive attributes has a stronger influence on bid values in the second round than in the third.

The results indicate that respondents have preferences for the introduced attributes both for bread and biscuits and that previous purchases of healthy foods as well as the realization of a connection between food and health are significant determinants for both products.

In the Sect. 4.2 we investigate if the reference point approach could shed more light on the process of consumer decision-making.

4.2 *Reference Point Effects*

The reference point effects approach suggests that each product is evaluated from a certain reference point, which can be represented by previously experienced prices or preferences. Closely related to the reference point is the idea of status quo bias in decision-making. Status quo is the current preferences of the individual that are preferred relative to new alternatives [24]. From a behavioral point of view this approach can be supported by, for example, causal model of gene technology acceptance, where acceptance is determined by perceived benefits and perceived risks [32]. It was also observed that when individuals face sequential choices, the subsequent decision is not independent from individuals initial status [29]. Thus, it can be assumed that in our experimental setting the subsequent bids of the participants are influenced by their own previous bids.

We fit the utility model that explains bid formation in the context of reference price effects on data from our second-price Vickrey auction. We estimate two models for revealed preferences. For the first model, a random utility approach as described in Sect. 4.1 is employed but now including additional parameters for reference point, gains and losses. Consequently, in the first case, the respondent maximizes utility when the perceived gains of an additional product attribute measured relative to a reference point lead to increasing their bid so that the respondent is able to purchase the product on the market. In our estimation including reference point effects we employ the same dependent variable as in the random utility estimation and the following specification:

$$V_{ij} = \alpha_j p + \beta_j z_i + \theta_j RP_{ij} + \gamma_{ij} Gain_{ij} + \lambda_{ij} Loss_{ij} + \varepsilon_{ij}, \qquad (3)$$

where V_{ij} is the deterministic part of the respondents i utility function; p is the average market price of the product; z_i is the individual i's characteristics; $RP_{ij}$is reference point for the product directly elicited from respondent i in the previous round; $Gain_{ij}$ is a dummy variable which equals 1 if auction bid exceeds reference price for this round, and equals 0 otherwise; $Loss_{ij}$ is a dummy variable which equals 1 if auction bid is less than reference price for this round, and equals 0 otherwise; and ε_{ij} is the error term.

For the second model we include probability weighting. The basis for our weighted model is Savages subjective utility model: $U_{ij} = p_{ij} V_{ij}$, where p_{ij} represents subjective probability [30]. We use subjective probabilities elicited from the stated choice model as weights following the behaviorist interpretation of subjective probabilities [4].

Incorporating probabilities into decision-making models builds on the assumption that decision outcomes are influenced by the importance of the outcomes to the decision maker. In our setting, we assume that the probabilities of purchasing foods with health and environmental benefits elicited from the survey reflect subjective probabilities that the respondent submits an auction bid higher than the market price of the product.

Using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as indicators of model fit, models involving reference point effects provide a significantly better fit than models based solely on the random utility approach. Moreover, models that use probability weights from a stated choice estimation have a better fit than estimations without weights (Table 6).

The results indicate that individual characteristics of respondents are mostly insignificant. Among the exeptions are previous purchases of healthy food (positive influence in the third round for biscuits); tradition (positive influence in the third round for biscuits); gender (positive influence for bread in the third round and negative for biscuits in the second round); and income (negatively influencing bids for biscuits in the second round). However, since this unsystematic evidence is hardly interpretable, we do not discuss it further.

Product-related characteristics like reference points and gains and losses are highly significant in all estimations. As we cannot interpret the coefficients in the

Table 6 Results of the logit estimations (1% significance level)

	2nd round		3rd round		2nd round, weights		3rd round, weights	
Utility function variables	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits
Factors related to product consumption								
Frequency	−0.28	−0.08	−0.53	−0.14	−0.22	––	−0.61	0.25
Price	−1.57	−0.92	1.67	0.18	−1.94	––	2.28	––
Taste	−0.61	−2.33	2.38	−0.14	−0.86	−2.00	2.58	−0.57
Healthiness	−0.18	0.19	2.67	−1.13	−0.43	0.72	3.13	−1.75
Novelty	−1.52	−2.50	1.57	−3.99	−1.61	−2.25	1.75	−3.80
Tradition	0.07	6.67	2.73	−1.53	−0.11	7.86***	2.94	−1.56
Connection between food and health	0.17	−1.48	1.60	0.21	−0.35	−4.31	1.11	0.06
Previous purchases of healthy food	0.19	0.02	−0.49	2.22***	0.04	−0.06	−0.22	2.79***
Consent to pay price premiums for products with healthy attributes	1.14	0.83	1.14	0.00	1.10	1.43	1.45	−0.07
Socio-demographic and lifestyle characteristics								
Gender	0.36	−1.77	2.29***	−0.16	0.22	−2.59***	1.89	−0.24
Income	−0.18	−0.85	0.33	−0.03	−0.06	−1.24***	0.39	−0.09
Nutr_ill	0.23	−0.15	1.00	0.79	0.12	−0.49	0.71	0.72
Sport	−0.05	−0.19	−0.09	−0.45	0.14	––	−0.15	−0.47
Smoke	0.77	0.78	0.64	−0.02	0.70	1.42	0.91	0.71
Alcohol consumption	0.57	−0.42	−0.93	0.75	0.46	−0.54	−1.36	0.49
Product-related characteristics								
Price	−0.18	−0.64***	−0.23	−0.42***	−0.14	−0.67***	−0.23	−0.44***
Reference point	0.27***	0.44***	0.48***	0.22***	0.24***	0.43***	0.49***	0.22
Gains	3.67***	4.43***	4.12***	2.82***	3.49***	4.75	4.60***	3.21***
Losses	−12.93***	−10.48***	−2.76	−1.35	−12.32***	−12.73	−2.44	−2.64
Constant	−3.87	9.26	−5.78	−0.40	−3.31	12.95	−4.66	−0.69
AIC	144.21	99.066	99.74	110.620	124.94	74.576	83.23	80.975
BIC	209.98	164.832	165.51	176.385	190.71	129.595	148.99	142.467
Pseudo R2	0.61	0.78	0.76	0.72	0.58	0.79	0.77	0.75

*, **, and *** denote significance at the .1, .05, .01 level, respectively. Standard errors are reported in parentheses

Table 7 Contrasts of predictive margins for statistically significant explanatory variables (covariates at observed values)

	2nd round		3rd round		2nd round, weighted		3rd round, weighted	
	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits	Bread	Biscuits
Average margin RP at gains	0.03***	0.02***	0.02***	0.01***	0.02***	0.02***	0.02***	0.01***
Average margin RP at losses	0.00	0.01***	0.02***	0.02***	0.00**	0.01	0.02***	0.01***
Gains	0.34***	0.22***	0.20***	0.17***	0.33***	0.22***	0.21***	0.17***
Losses	−0.61***	−0.42***	−0.12*	−0.08	−0.58***	−0.50***	−0.10*	−0.14

*, **, and *** denote significance at the .1, .05, .01 level, respectively. Standard errors are reported in parentheses

estimation directly, we look at the predicted probabilities. Results regarding average predicted probabilities for reference points at gains and losses indicate the response of the dependent variable, in our case the willingness to purchase product with health and environmental attributes introduced, to the reference point when attributes are being perceived either as gains or as losses. Due to the fact that a reference point is represented by a continuous variable, the values in the table indicate only the rate at which dependent variable would be changing if this rate is constant. Nonetheless, it indicates that there is a relatively low influence of reference point on the purchase decision (Table 7).

Predicted probabilities for gains and losses demonstrate that perceived gains of the product attributes increase the probability of purchase by the range of 17–34 % depending on the estimation method. More importantly, perceived losses decrease the probability of purchase by the range of 8–61 % depending on the estimation method that fit the idea of an asymmetrical response to gains and losses. Moreover, graphs of contrasts of probabilities show that lower reference points are more influenced by both gains and losses for both products. Lower auction bids are submitted by consumers who are either limited by budget or do not have salient preferences for products with health and environmental benefits. In any case, their decisions are more influenced by the attributes introduced. More prominent asymmetrical effect can be observed for the second auction round compared to the third. The combination of two attributes in the third round reduces asymmetrical effect, thus indicating the importance of attributes introduced (Figs. 1 and 2).

5 Conclusions and Discussion

In this research we inferred additional information on consumer preferences for foods with health and environmental attributes by applying both traditional random utility and reference point approaches to experimental auction data. A random utility

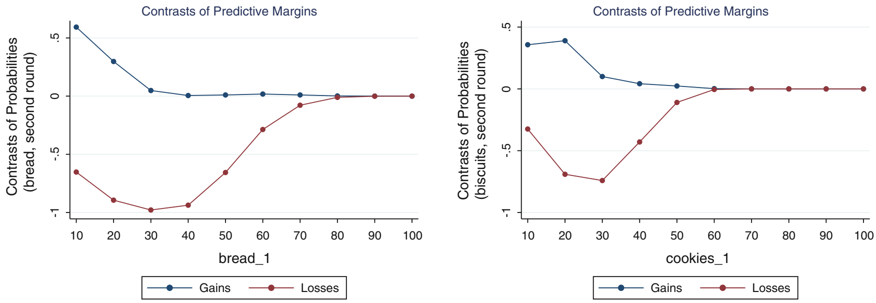


Fig. 1 Contrasts of predictive margins for gains and losses for bread and biscuits in the 2nd round of the auction

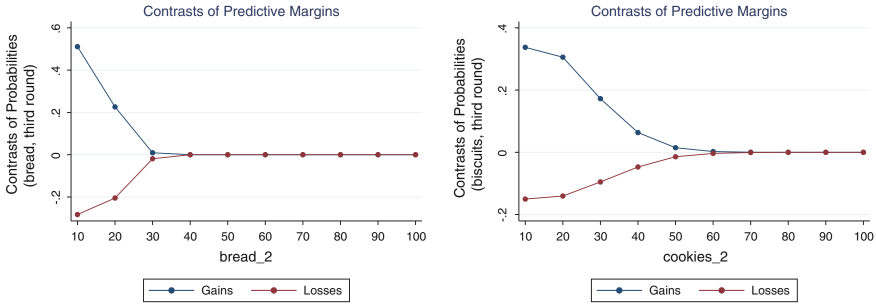


Fig. 2 Contrasts of predictive margins for gains and losses for bread and biscuits in the 3rd round of the auction

approach allowed us to identify the most significant factors influencing the bids. Differences between utility levels inferred from hypothetical and revealed preferences indicated significant overestimation in consumer stated choices, which is in line with previous research [20].

Differences between preferences underlying utility values in stated and revealed logit can have different reasons. However, it seems reasonable to assume that framing of the choices played a significant role in our setting. The survey that was presented to the respondents before the auction was aimed at eliciting attitudes, knowledge and preferences, while the auction procedure was strictly aimed at valuation. Consequently, respondents might have perceived these two parts of the experiment as separate, which resulted in different factors influencing stated and revealed choices.

A random utility approach provides value measures in absolute terms, whereas the reference point effects approach is based on relative measures according to a specific reference point. The results of the reference point approach indicate the presence of reference points and asymmetrical effects of gains and losses in consumer preferences. Deviation from reference point might explain preference formation during the auction. Consumers value health and environmental attributes when compared to the average market price or bid in the previous round. However, when compared to the reference point both the healthy and the environmental attribute produce not only gains but also losses, and these perceived losses can significantly decrease the probability of purchase.

When participants demonstrate loss aversion towards novel products, it is necessary to pay more attention to the factors that comprise possible losses for participants. In other words, if losses outweigh gains, then factors that can possibly be perceived as losses perhaps require even more attention than gains. In the context of bioeconomy losses might include the broad range of concerns related to biotechnology and under closer investigation lead to deeper understanding of consumer resistance.

Moreover, recent research demonstrates that for regularly purchased products (like bread) consumers in their valuations rely heavily on the prices observed on the market [23]. Consumers do not rely on subjective utility that can be derived from the use of novel product but assume the value of the product based on market experience. It can explain difference in valuing the same attributes of different products, in our case differences between bread and biscuits. This approach could also shed more light on differences between Moscow and Irkutsk markets, however, due to limited amount of observations, we could not perform this analysis.

Reference point model fit measured by PseudoR^2 and Akaike Information Criterion and Bayesian Information Criterion indicate that part of consumer-specific heterogeneity was explained in the models that include reference points. Models taking into account weights, where weights are formed from previous beliefs fit better than unweighted models.

The auction data presented in this chapter included sequential bidding. The sequential order of introducing product attributes might have influenced the outcomes. Loewenstein and Prelec [17] demonstrated that when introduced to the sequence of outcomes people show preference for improvement, which in our case might have inflated the stakes, as we have introduced only positive product characteristics. This drawback could be overcome in the future research by adjusting the experimental design.

References

1. Adamowicz, W., Louviere, J., Williams, M.: Combining revealed and stated preference methods for valuing environmental amenities. J. Envi. Econ. Manag. **26**, 27192 (1994)
2. Aseff, J.G.: Learning to play second-price auctions, an experimental study. Econ. Lett. **85**(2), 279286 (2004)
3. Banerji, A., Gupta, N.: Detection, identification, and estimation of loss aversion: evidence from an auction experiment. Am. Econ. J. Microecon. **6**(1), 91133 (2014)
4. Baron, J., Frisch, D.: Ambiguous probabilities and the paradoxes of expected utility. In: Wright, G., Ayton, P. (eds.) Subjective Probability, pp. 1–20. Wiley, Chichester, Sussex (1994)
5. Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R.: A test of the theory of reference-dependent preferences. Quat. J. Econ. (1997). May
6. Bell, D.R., Lattin, J.M.: Looking for loss aversion in scanner panel data: the confounding effect of price response heterogeneity. Mark.Sci. **19**(2), 185200 (2000)

7. Cheng, L., Monroe, K.: An appraisal of behavioral price research (part 1): price as a physical stimulus. AMS Rev. **3**, 103–129 (2013)
8. Frewer, L., Scholderer, J., Lambert, N.: Consumer acceptance of functional foods: issues for the future. Br. Food J. **105**(10), 714731 (2003)
9. Green, W.: Econometric Analysis, 5th edn. Prentice Hall, Upper Saddle River, N.J (2003)
10. Hardie, B., Johnson, E.G., Fader, P.S.: Modeling loss aversion and reference dependence effects on brand choice. Mark Sci. **12**(4), 378394 (1993)
11. Hess, S., Rose, G.M., Hensher, D.A.: Asymmetric preference formation in willingess to pay estimates in discrete choice models. Transp. Res. Part E Logist. Transp. Rev. **44**(5), 847863 (2008)
12. Hu, W., Adamowicz, W., Veeman, M.M.: Labeling context and reference point effects in models of food attribute demand. Am. J. Agric. Econ. **88**(4), 10341049 (2006)
13. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**(2), 263292 (1979)
14. Kalwani, M.U., Yim, C.K.: Consumer price and promotion expectations: an experimental study. J. Mark Res. **29**(February), 90100 (1992)
15. Lattin, J., Bucklin, R.: Reference effects of price and promotion on brand choice behavior. J. Mark Res. **26**(3), 299310 (1989)
16. Lee, A.: A closer look at reference price: a commentary. AMS Rev. **3**(3), 151–154 (2013)
17. Loewenstein, G.F., Prelec, D.: Preferences for sequences of outcomes. Psychol. Rev. **100**(1), 91108 (1993)
18. Lusk, J.L., Feldkamp, T., Schroeder, T.C.: Experimental auction procedure: impact on valuation of quality differenciated goods. Am. J. Agric. Econ. **86**(2), 389405 (2004)
19. Lusk, J.L., Schroeder, T.C.: Auction bids and shopping choices. Adv. Econ. Anal. Policy **6**(1), 137 (2006)
20. Lusk, J.L., Shogren, J.F.: Experimental Auctions. Methods and Applications in Economic and Marketing Research. Cambridge University Press, New York (2007)
21. Marette, S., Roosen, J., Blanchemanche, S., Feinblatt-Meleze, E.: Functional food, uncertainty and consumers choices: a lab experiment with enriched yoghurts for lowering cholesterol. Food Policy **35**(5), 419428 (2010)
22. Marette, S., Roosen, J., Blanchemanche, S., Verger, P.: The choice of fish species: an experiment measuring the impact of risk and benefit information. J. Agric. Resour. Econ. **33**(1), 118 (2008)
23. Mazar, N., Koszegi, B., Ariely, D.: True context-dependent preferences? the causes of market-dependent valuations. J. Behav. Dec. Mak. **27**, 200–208 (2014)
24. McFadden, D.: Rationality for economists? J. Risk Uncertain. **19**(1–3), 73105 (1999)
25. Noussair, C., Robin, S., Ruffieux, B.: Revealing consumers willingness-to-pay: a comparison of the bdm mechanism and the vickrey auction. J. Econ. Psychol. **25**(6), 725741 (2004)
26. Onwezen, M.C., Bartels, J.: Which perceived characteristics make product innovations appealing to the consumer? a study on the acceptance of fruit innovations using cross-cultural consumer segmentation. Appetite **57**(1), 5058 (2011)
27. Putler, D.: Incorporating reference price effects into a theory of consumer choice. Mark Sci. **11**(3), 287309 (1992)
28. Rozan, A., Stenger, A., Willinger, M.: Willingness-to-pay for food safety: an experimental investigation of quality certification on bidding behaviour. Eur. Rev. Agric. Econ. **31**(4), 409425 (2004)
29. Samuelson, W., Zeckhauser, R.: Status quo bias in decision making. J. Risk Uncertain. **1**, 759 (1988)
30. Savage, L.J.: The Foundations of Statistics, 2nd edn. Dover Publications Inc., New York (1972)
31. Shogren, J.F., et al.: Auction mechanisms and the measurement of WTP and WTA. Resour. Energy Econ. **23**(2), 97109 (2001)
32. Siegrist, M.: The influence of trust and perceptions of risks and benefits on the acceptance of gene technology. Risk Anal. **20**(2), 195203 (2000)
33. Siro, I., et al.: Functional food product development, marketing and consumer acceptance-a review. Appetite **51**(3), 456467 (2008)

34. Tversky, A., Kahneman, D.: Rational choice and the framing of decisions. J. Bus. **59**(4), 251278 (1986)
35. Tversky, A., Kahneman, D.: Loss aversion in riskless choice: a reference-dependent model. Q. J. Econ. **106**(4), 10391061 (1991)
36. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. J. Risk Uncertain. **5**(4), 297323 (1992)
37. Umberger, W.J., Feuz, D.M.: The usefulness of experimental auctions in determining consumers willingness-to-pay for quality-differentiated products. Rev. Agric. Econ. **26**(2), 170185 (2004)
38. Verbeke, W.: Functional foods: consumer willingness to compromise on taste for health? Food Qual. Prefer. **17**(1–2), 126131 (2006)
39. Vickrey, W.: Counterspeculation, auctions, and competitive sealed tenders. J Financ. **16**(1), 8–37 (1961)

Corporate Asset Pricing Models and Debt Contracts

Martin Dózsa and Karel Janda

Abstract Our contribution aims to provide an introduction to the theory of corporate asset pricing models and explain the potential of their usage in the design of credit contracts. We describe the evolution of structural models starting from the basic Mertonian framework through the introduction of a default barrier, and ending with stochastic interest rate environment. Further, with the use of game theory analysis, the parameters of an optimal capital structure and safety covenants are examined. Furthermore an EBIT-based structural model is introduced that considers stochastic default barrier. Such set-up is able to catch the different optimal capital structures in various business cycle periods, as well as bankruptcy decisions dependent on the state of the economy. The effects of an exogenous change in the risk-free interest rate on the asset value, probability of default, and optimal debt ratio are also explained.

Keywords Credit contracts · Stochastic default barrier · Asset pricing · EBIT-based models · Structural models

1 Introduction

In the past decades financial markets rapidly gained on complexity due to an increased demand for risk diversification and hedging. A number of sophisticated instruments was developed that capture various aspects of price movements, correlations of assets, macro-economical developments, and other changes that might affect the future

M. Dózsa · K. Janda (✉)
Institute of Economic Studies, Faculty of Social Sciences, Charles University
in Prague, Prague, Czech Republic
e-mail: Karel-Janda@seznam.cz

M. Dózsa
e-mail: martin@dozsa.cz

K. Janda
Department of Banking and Insurance, Faculty of Finance and Accounting,
University of Economics, Prague, Czech Republic

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_11

income generated by the considered securities. The pricing of these securities was not sufficiently accurate using the traditional asset pricing models. In the search for new methods two different approaches appeared. One stream of literature (called the reduced-form approach) focused on finding a purely mathematical way of asset pricing, without the effort of finding any economical intuition behind the models. In contrast, the other group of academics studied the firm and its evolution. These, so-called structural models have an intuitive connection to the underlying economics, and therefore they can be helpful in understanding the reasons of price movements.

This work fits in the category of structural approaches. First it gives a brief overview to the development of these models, and proposes their extension to a stochastic interest rate environment. Second, it uses these models to examine the effects of parameter settings in debt contracts, and therefore gives a guidance for the design of an optimal credit contract that maximizes firm value. With the introduction of a stochastic interest rate environment, it is possible to consider the implications of the business cycle period on the optimal debt ratio, and—using stochastic default barrier—on the bankruptcy decision as well. Game theory is also invoked, therefore agency costs arising from asymmetric information are predicted and minimized with the help of safety covenants and properly chosen parameters.

2 Asset Pricing Models

Due to the risk-averse human nature the price of an asset is dependent on its riskiness (that is, on the volatility of its future returns): investors price assets below their expected payoff if they bear some risk. However, the idea of a risk-neutral probability measure deals with this issue: it is possible to adjust the probabilities of future states for risk in a way that assets are priced at their expected values.[1] To derive this risk-neutral probability measure we need the assumption that market prices include all available information, since known fair prices are needed in order to create a measure that produces expected values equal to these fair prices. Furthermore this risk-neutral probability measure is unique if markets are complete.

Models that require the assumption that market prices incorporate all available information are called market information based models. They can be further divided to structural and reduced-form models.

Models representing the first category are based on the Merton [42] framework that employs the option pricing theory presented by Black and Scholes [9]. In Merton's work a company defaults at the maturity of its debt if the value of its assets is below the sum of its liabilities. Default prior maturity is not possible. The subsequent models relaxed this assumption as well as others taken by Merton. The common attribute of these models is that they concentrate on the structural characteristics of a company, including asset volatility and financial leverage.

[1] The probability measure that reflects the true probabilities is called the physical measure.

By contrast, reduced-form (aka hazard rate) models ignore structural characteristics, and treat bankruptcy as a possible exogenous event that is described as the first jump time of a point process, without trying to explain the reason of default. This approach was first proposed by Jarrow and Turnbull [30] and later extended in several works, for example [29, 37] or [16].

2.1 Merton's Structural Model

In his pathbreaking paper, Merton [42] paralleled the value of equity in a leveraged firm to a European call option on the firm's assets and used the option pricing theory developed by Black and Scholes [9] to value it. A corresponding debt is a zero-coupon bond with finite maturity with a promised terminal payoff B. This rather simplified description has many unrealistic restrictions, however, because of its simplicity and new perspective Merton built the basics of the framework used in structural models.

A large and growing body of literature has relaxed one or more assumptions posed by Merton. Some of the most important extensions are: more complex capital structure and safety covenants [8], interest paying debt [20], bankruptcy costs and tax benefits [34], short and long term debt types [53], or stochastic interest rate [11, 14, 23, 36].

The original framework's assumptions, mainly coming from the Black and Scholes [9] option pricing theory are[2]:

(A.1) There are no transactions costs, taxes, or problems with indivisibilities of assets.

(A.2) There is a sufficient number of investors with comparable wealth levels so that each investor believes that he can buy and sell as much of an asset as he wants at the market price.

(A.3) There exists an exchange market for borrowing and lending at the same rate of interest.

(A.4) Short-sales of all assets, with full use of the proceeds, are allowed.

(A.5) Trading in assets takes place continuously in time.

(A.6) The Modigliani-Miller theorem that the value of the firm is invariant to its capital structure obtains.

(A.7) The Term-Structure is "flat" and known with certainty. I.e., the price of a riskless discount bond which promises a payment of one dollar at time τ in the future is $P(\tau) = e^{-r\tau}$ where r is the (instantaneous) riskless rate of interest, the same for all time.

(A.8) The dynamics for the value of the firm, V, through time can be described by a diffusion-type stochastic process with Stochastic Differential Equation (SDE)[3]

[2] The assumptions are written exactly in a way as Merton wrote them, except for the symbols used.

[3] This process is called Geometric Brownian Motion.

$$dV = (\mu V - C)dt + \sigma V dW, \tag{1}$$

> where μ is the instantaneous expected rate of return on the firm per unit time, C is the total dollar payout by the firm per unit time to either its shareholders or liability-holders (e.g., dividends or interest payments) if positive, and it is the net dollars received by the firm from new financing if negative; σ^2 is the instantaneous variance of the return on the firm per unit time; dW is a standard Gauss-Wiener process.

Suppose a security with market value, Y dependent on the value of a firm. More specifically, its price can be written as a function of the firm value V, and time t: $Y = F(V, t)$. The dynamics of this security can be formally written using a SDE as

$$dY = [\mu_Y Y - C_Y]dt + \sigma_Y Y dW_Y, \tag{2}$$

where μ_Y, C_Y, σ_y and W_Y and defined similarly as in (1). Using the stochastic equivalent of chain-rule, the so-called Itō's Lemma we also have:

$$\begin{aligned} dY &= F_V dV + \frac{1}{2} F_{VV}(dV)^2 + F_t \\ &= \left[\frac{1}{2}\sigma^2 V^2 F_{VV} + (\mu V - C)F_V + F_t\right] dt + \sigma V F_V dW, \end{aligned} \tag{3}$$

where subscripts denote partial derivatives, and the second equation comes from (1). Comparing terms in (2) and (3) we have

$$\mu_Y Y \equiv \frac{1}{2}\sigma^2 V^2 F_{VV} + (\mu V - C)F_V + F_t + C_Y \tag{4}$$

$$\sigma_Y Y \equiv \sigma V \tag{5}$$

$$dW_Y \equiv dW \tag{6}$$

The last equation indicates that Y_t and V_t are perfectly correlated, as they are driven by the same stochastic parameter. This implies the existence of such linear combination of these securities that the resulting payoff is non-stochastic. Using this fact Merton constructed a portfolio of three securities V, Y and riskless debt in a way that the initial investment was zero.[4] He showed that any security Y whose value can be written as a function of the firm value and time has to satisfy the following equation:

$$0 = \frac{1}{2}\sigma^2 V^2 F_{VV} + (\mu V - C)F_V - rF + F_t + C_Y \tag{7}$$

As we can see, F depends on the value of the firm, time, interest rate, the volatility of the firm's value, the payout policy of the firm and the payout policy to the holders

[4] For the details about the construction of this portfolio, and for the complete derivation of Eq. (7) see [42] pp. 451–452.

of Y. It does not depend on the expected rate of return neither the risk preference of the investors. This is the result where the idea of risk-neutral valuation comes from. Also it should be noted, that the only thing that distinguishes one security from the other (debt vs. equity) is a pair of boundary conditions.

For pricing a simple corporate bond Merton took four further assumptions:

(A.9) The corporation has two classes of claims, a single homogeneous class of debt and the residual claim, called equity.

(A.10) The firm commits to pay $B to the bondholders at date T.

(A.11) If the payment is not met at T, the bondholders immediately take over the company, and so the shareholders receive nothing.

(A.12) The firm cannot issue any new claims that are not junior to the original one nor can pay dividends or do share repurchase before T.

As it can be seen this set-up ensures no default prior to maturity. Using Eq. (7) for the value of the debt, D, setting $C = C_Y = 0$ in line with the assumptions and defining $\tau = T - t$, so thus $D_t = -D_\tau$ we can write

$$0 = \frac{1}{2}\sigma^2 V^2 D_{VV} + rVD_V - rD - D_\tau \tag{8}$$

Denoting the value of equity as E and using (1), we have $V = D(V, \tau) + E(V, \tau)$. As E and D are non-negative, we know:

$$D(0, \tau) = E(0, \tau) = 0$$

and also $D(V, \tau) \leq V$, that is for $V > 0$ we have the other boundary condition

$$D(V, \tau)/V \leq 1$$

As the payment is made exactly when $V(T) > B$, the initial condition for the debt at $\tau = 0$ is

$$D(V, 0) = \min[V, B]$$

The function $D(V, \tau)$ can be found using (8) and the above boundary conditions using standard methods as separation of variables. However, as Merton noticed, the problem can be transformed to another, already solved. For the value of equity holds $E(V, \tau) = V - D(V, \tau)$, so the solution for equity is given by (7):

$$0 = \frac{1}{2}\sigma^2 V^2 E_{VV} + rVE_V - rE - E_\tau \tag{9}$$

with a corresponding initial condition

$$E(V, 0) = \max[0, V - B]$$

and the boundary conditions $E(0, \tau) = 0$ and $E(V, \tau)/V \leq 1$. This is identical to the equations for an European call option on a non-dividend-paying stock in the Black-Scholes option pricing model. The firm value corresponds to the stock price, the equity to the option value and B to the exercise price.

Therefore the equity price is

$$E(V, \tau) = V\Phi(d_1) - Be^{-r\tau}\Phi(d_2), \tag{10}$$

where

$$\begin{aligned} d_1 &= \frac{\ln(V/B) + \left(r + \frac{1}{2}\sigma^2\right)\tau}{\sigma\sqrt{\tau}} \\ d_2 &= d_1 - \sigma\sqrt{\tau} \end{aligned}$$

and $\Phi(\cdot)$ is the cumulative standard normal distribution.

As $D = V - E$, the debt value can be expressed as

$$D(V, \tau) = V\Phi(-d_1) + Be^{-r\tau}\Phi(d_2) \tag{11}$$

with the d_1 and d_2 as in (10).

2.2 First Passage Time Approach

The original Merton [42] model described in the previous section uses several assumptions that limit its practical implementability. One of the most unrealistic restrictions is the impossibility of default before maturity. To solve this problem Black and Cox [8] came with a set-up where default occurs if the firm value touches a threshold level. This level is called the Default Barrier (DB), and generally can be constant [34, 36], deterministic [8, 35] or stochastic [11, 14] function of time. Models with a DB not only explain early default, but are also able to produce a large variety of Recovery Rates (RRs) and therefore reflect more precisely factors as bond covenants, bankruptcy costs or taxes.

The name of the First Passage Time (FPT) models corresponds to the method how the default is described mathematically: since the evolution of the firm value is represented using a Geometric Brownian Motion (GBM), it is possible to transform the probability distribution of the default to the FPT of a Wiener process. These models can be also divided to two groups in dependence on the determination of the DB: it can be set exogenously [8, 36], or be an endogenous result of an optimization process [34, 57]. The notation used throughout the section follows the one introduced in the description of Merton's model, unless it is explicitly defined otherwise.

2.2.1 Black and Cox Model

Black and Cox [8] extended the original Merton [42] framework to include several features of debt contracts, namely safety covenants, subordinated bonds, and restriction on asset sales. Since this section discusses basic asset pricing methods, only the introduction of a DB will be described.[5]

The evolution of the firm value is the same as in the Merton model [42], except a restriction that the continuous dividend payment received by the stockholders is a constant fraction of the firm value. Therefore Eq. (1) takes the form

$$dV = V(\mu - c)dt + \sigma V dW \tag{12}$$

with a constant $c = C/V$ representing the payout ratio received by the equity holders. Again, the short-term interest rate is assumed to be constant, and so the interest-rate risk is disregarded. The original case described in [8] also assumes zero bankruptcy costs.

A safety covenant, that provides a right for the bondholders to force bankruptcy if the firm is performing poorly, is introduced. This poor performance is signalled by the fall of the firm value under a time-dependent default barrier defined as $\bar{v}(t) = Ke^{-\gamma(T-t)}$, $t \in [0, T)$ for some constants $K > 0$ and γ. The creditors take over the firm as soon as the firm value hits this barrier. Consequently default could be triggered in two ways: prior to maturity (by reaching the threshold level) or at maturity, if the firm value was above the DB but is below the debt principal at T. To simplify the notation let us set the default barrier as one function:

$$v_t = \begin{cases} \bar{v}(t) & \text{for } t < T, \\ B & \text{for } t = T. \end{cases}$$

The default time τ is

$$\tau = \inf \{t \in [0, T] : V_t < v_t\}.$$

We also assume the following:

$$V_0 > \bar{v}(0)$$

$$Ke^{-\gamma(T-t)} \leq Be^{-r(T-t)}, \quad \forall t \in [0, T]$$

i.e. the firm is not in default initially and the default barrier (and hence the payment to the bondholder) is never higher than the present value of the principal amount. This holds also for $t = T$, therefore $K \leq L$.

[5] For pricing of more complex capital structures and the issue of contractural design see the original work of [8].

Zero-Coupon Bond In Merton's model the debt pricing function solved Eq. 8. The analogous Partial Differential Equation (PDE) for zero-coupon debt value with default barrier is

$$0 = \frac{1}{2}\sigma^2 V^2 D_{VV} + (r-c)VD_V - rD + D_t \tag{13}$$

with the boundary condition

$$D(Ke^{-\gamma(T-t)}, t) = Ke^{-\gamma(T-t)}$$

and terminal condition

$$D(V,T) = \min(V,B).$$

Equation (13) can be solved using the classical methods used for PDEs or with a probabilistic approach.[6]

Note, that similarly as the equity value in [42] corresponds to a call option, it corresponds to a down-and-out barrier option here. Using the in-out parity (i.e. the plain vanilla option price equals to the sum of down-and-out and down-and-in barrier options price, all having the same strike price, underlining asset, maturity and the last two having the same barrier as well), the equity has a lower value by the price of a down-and-in barrier option in the presence of a DB. As there are no bankruptcy cots, this value is transferred to the bondholder.

Perpetual Coupon Bond A perpetual coupon bond has infinite maturity and continuous coupon payment at a constant rate c_D.[7] The net cost of the coupon is financed by issuing additional equity. Its price $D_{c_D}(t)$ equals

$$D_{c_D}(t) = \lim_{T\to\infty} E\left(\int_t^T c_D e^{-r(s-t)} 1_{\{s<\bar{\tau}\}} ds\right) + \lim_{T\to\infty} E\left(Ke^{\gamma(\bar{\tau}-T)} e^{-r(\bar{\tau}-t)} 1_{\{t<\bar{\tau}<T\}}\right)$$

under risk-neutral probability measure with 1 used as a symbol for indicator function. Since the coupon payments are constant it is straightforward to define the default barrier constant as well, i.e. set $\gamma = 0$. With the assumption that dividends paid to equity holders are zero (that is $c = 0$) D_{c_D} can be written as[8]

$$D_{c_D} = \frac{c_D}{r}\left(1 - \left(\frac{\bar{v}}{V_t}\right)^{\alpha}\right) + \bar{v}\left(\frac{\bar{v}}{V_t}\right)^{\alpha}, \tag{14}$$

with $\alpha = 2r/\sigma^2$.

[6] The solution of (13) can be found in [8] p. 356.

[7] Here we use the subscript D in order to distinguish this pay-out from c, which was the payout ratio to equity holders.

[8] For the mathematical derivation see [7] p. 81 and the preceding calculations.

2.2.2 Leland's Model

Leland [34] extended the perpetual coupon bond model described above by incorporating bankruptcy costs and tax benefits. Now V is a variable for the "asset value" of the firm; the total firm value is V less the expected costs of bankruptcy plus the value of the tax shield. V follows the same diffusion process as in (12) with no dividend payments ($c = 0$):

$$dV = V\mu dt + \sigma V dW,$$

hence V is not affected by the financial structure of the firm, thus the difference between coupon payments and tax benefits is financed by equity dilution.

When bankruptcy occurs at level $V_t = V_B$ a fraction $0 \leq \omega \leq 1$ is lost as costs due to bankruptcy, and the debt holders receive the remaining $(1-\omega)V_B$ leaving the equity holders with nothing. The value of the bond can be written as

$$D_{c_D}(V_t) = \frac{c_D}{r}\left(1 - \left(\frac{\bar{v}}{V_t}\right)^{\alpha}\right) + (1-\omega)\bar{v}\left(\frac{\bar{v}}{V_t}\right)^{\alpha}. \tag{15}$$

Note that with $\omega = 0$ this is identical to (14). If we denote $p_B = (\bar{v}/V_t)^{\alpha}$ (15) becomes

$$D_{c_D}(V_t) = \frac{c_D}{r}(1-p_B) + (1-\omega)\bar{v}p_B.$$

p_B represents the value of a contingent claim that pays \$1 when bankruptcy occurs, $\omega\bar{v}p_B$ is the present value of expected bankruptcy costs, and $c_D/r\,(1-p_B)$ is the present value of expected coupon payments. Consequently the value of the tax benefits is equal to:

$$TS = T_c\frac{c_D}{r}(1-p_B),$$

where T_c is the corporate tax rate.

The total value of the firm, denoted by $G(V_t)$ is therefore equal to

$$G(V_t) = V_t - \omega\cdot\bar{v}\cdot p_B + T_c\frac{c_D}{r}(1-p_B).$$

Since the total value of the firm is equal to the sum of its equity and debt value, the shareholders' claim can be found as

$$E(V_t) = G(V_t) - D_{c_D}(V_t)$$
$$E(V_t) = V_t - (1-T_c)\frac{c_D}{r}(1-p_B) - \bar{v}\cdot p_B.$$

Intuitively the value of equity is equal to the value of firm's assets less the present value of expected coupon payments reduced by tax and the contingent claim on $\bar{v}$. Note that the value of equity is not dependent on the bankruptcy costs, and so that is paid in full by the bondholders.

2.3 Models with Stochastic Interest Rates

One of the shortcomings of the Black and Cox [8] model is the assumption of constant and known risk-free interest rate. This restriction is relaxed in models with stochastic interest rates. Because our work[9] assumes stochastic interest rate as well, we will make a review of the relevant literature at this point.

2.3.1 Longstaff and Schwartz

Longstaff and Schwartz [36] price corporate bonds reflecting both interest rate risk and credit risk using risk-neutral probability measure for both stochastic processes. The evolution of the short-term interest rate is inherited from the Vasicek [52] model:

$$dr_t = (a - br_t)dt + \sigma_r d\tilde{W}_t,$$

and the firm-s value is driven by the

$$dV_t = V_t(r_t dt + \sigma_V dW_t^*)$$

SDE. As we can see the constant drift from the Leland [34] model is replaced by the stochastically evolving short-term interest. Furthermore, following Longstaff and Schwartz [36] we have the following properties:

- Brownian motions $\tilde{W}$ and W^* are correlated with the instantaneous correlation $\rho_{V,r}$.
- DB is represented as a constant threshold level $\bar{v}$.
- Recovery Rate (RR) is independent on the default time, proportional to the face value of the bond and paid out at maturity.
- $\bar{v} \geq B$, hence the debt is repaid in full if default does not occur prior maturity.[10]
- The firm has one or more debt classes with different recovery rates $(1 - \omega_i)$, where ω_i is the writedown rate for the ith class. The seniority of the claims is already reflected in the writedown rates, and therefore does not play essential role.[11] It is natural to suppose the following relationship: $\bar{v} = \sum_{i=1}^{k}(1 - \omega_i)B_i$ with B_i $(\sum_{i=1}^{k} B_i = B)$ representing the total face value of debt from the ith class.

If we define τ, the time of default in the traditional way, that is

$$\tau - \inf\{t \in [0, T] : V_t < \bar{v}\},$$

[9] See Sect. 4.

[10] In fact this inequality is not explicitly wrote down by [36], however it is implicitly assumed.

[11] Note that this set-up can easily catch Absolute Priority Rule (APR) violations.

then the bond's payoff at T can be written as

$$D_i(V_T, T) = B(1 - \omega_i 1_{\{\tau \le T\}}).$$

For finding an analytic solution of the bond value at time $t < T$ with given V_t there are basically two ways: by solving the fundamental PDE with the corresponding boundary and terminal conditions, or alternatively, by probabilistic approach. A closed-form solution however has not yet been produced using any of them. For this reason—even if some quasi-explicit results can be obtained analytically—numerical computations are required in order to obtain the results of the model. Such computations were made by the authors as well as others [14, 33]. A shortcoming of this model is, that it produces credit spreads close to zero for low debt maturities.

2.3.2 Briys and de Varenne

Briys and de Varenne [11] submitted a model that addressed some restrictive features and assumptions of the then available literature. For example, the previously analyzed Longstaff and Schwartz [36] model cannot work with a default barrier that would be lower than the present value of the debt principal. Their work also assumes stochastic default barrier, as it is derived from the instantaneous short-term interest.

The short-term rate dynamics follows the so-called generalized Vasicek model, which is a mean-reverting stochastic function:

$$dr_t = a(t)(b(t) - r_t)dt + \sigma(t)d\tilde{W}_t,$$

where $a, b, \sigma : [0, T] \to \mathbb{R}$ are known, deterministic functions. Consequently the price of a default-free zero-coupon bond, P follows the dynamics

$$dP(t, T) = P(t, T)(r_t dt + b(t, T)d\tilde{W}_t)$$

for some deterministic $b(\cdot, T) : [0, T] \to \mathbb{R}$. The firm value V is assumed to follow the process

$$\frac{dV_t}{V_t} = r_t dt + \sigma_V(\rho d\tilde{W}_t + \sqrt{1 - \rho^2} d\hat{W}_t),$$

with constant $\sigma_V > 0$, and mutually independent Brownian motions $\tilde{W}$ and $\hat{W}$. The local correlation coefficient between the risk-free rate and firm value is $\rho = \rho_{V,r}$. If we denote $W^* = \rho d\tilde{W}_t + \sqrt{1 - \rho^2} d\hat{W}_t$, it is visible that the firm value process is defined in the same fashion as in [34].

The DB is defined as the price of a default-free bond with the same maturity and some face value $K \in (0; B]$ not greater than the defaultable bond principal:

$$v_t = \begin{cases} K \cdot P(t, T) & \text{for } t < T, \\ B & \text{for } t = T. \end{cases}$$

The default time is, as usually,

$$\tau = \inf\{t \in [0, T] : V_t < v_t\}.$$

The payoff at default is dependent on τ: for $\tau < T$ the bondholders receive a $(1 - \omega_2)$ part of the remaining assets, whereas for $\tau = T$ this payoff ratio may be different, and is represented as $(1 - \omega_1)$. The remaining ω_1 respectively ω_2 part is lost as bankruptcy cost and/or paid out to equity holders (APR). The bond's final cash flow at T is therefore

$$D(V_t, T) = (1 - \omega_2)B1_{\{\tau<T\}} + (1 - \omega_1)V_T 1_{\{\tau=T\}} + B1_{\{\tau>T\}}$$

If the bond price volatility function $b(t, T)$ is known, than the price of a defaultable corporate bond can be derived as a closed-from solution:

$$D(t, T) = P(t, T) \cdot [B - D_1 + D_2 - \omega_2 R_2 - \omega_1 R_1], \tag{16}$$

where $F_t = V_t/P(t, T)$

$$\begin{aligned}
D_1 &= B\Phi(d_1) - F_t\Phi(d_2),\\
D_2 &= K\Phi(d_5) - (F_t L/K)\Phi(d_6),\\
R_2 &= F_t\Phi(d_4) + K\Phi(d_3),\\
R_1 &= F_t\big(\Phi(d_2) - \Phi(d_4)\big) + K\big(\Phi(d_5) - \Phi(d_3)\big),
\end{aligned}$$

with

$$\begin{aligned}
d_1 &= \frac{\ln(B/F_t) + \frac{1}{2}\sigma^2(t, T)}{\sigma(t, T)} = d_2 + \sigma(t, T),\\
d_3 &= \frac{\ln(K/F_t) + \frac{1}{2}\sigma^2(t, T)}{\sigma(t, T)} = d_4 + \sigma(t, T),\\
d_5 &= \frac{\ln(K^2/(F_t B)) + \frac{1}{2}\sigma^2(t, T)}{\sigma(t, T)} = d_6 + \sigma(t, T),
\end{aligned}$$

and

$$\sigma^2(t, T) = \int_t^T \big((\rho\sigma_V - b(u, T))^2 + (1 - \rho^2)\sigma_V^2\big)\, du.$$

Let us analyze (16) here: $B - D_1$ corresponds to the Mertonian valuation (i.e. risk-free bond less put-to-default option), D_2 is associated with the value brought to the debt holders by the possibility of early default triggered by safety covenant. The last two terms, $\omega_2 R_2$ and $\omega_1 R_1$, are both positive,[12] and represent the costs of early

[12] See [7] pp. 105–106.

default and default at maturity respectively. It is therefore clear that the bond's price is decreasing in ω_1 and ω_2.

3 Credit Contracts

This section explains the reasons for issuing debt, and gives an insight to the design of credit contracts that aims for the maximization of firm value and the prevention of unexpected losses in the contracting parties' claims. The answer to this problem is given using the tools described in the previous section, where we briefly introduced theoretical works that help us in pricing the two basic types of claims on the firm's assets: debt and equity.

3.1 Capital Structure

The capital structure of a firm refers to the proportion of securities that ensure the needed funds for financing the firm's projects. These securities have two basic types: a riskier asset called equity and a relatively safe one, the debt. Equity has two further sub-groups (preferred and common), debt has many flavours, and furthermore there exists a group called "hybrid securities" including, for example convertible bonds. In this work we will concentrate on the two basic types only, however the model presented in Sect. 4 can be easily extended to more complex capital structures as well.

The value of the firm is therefore the sum of the market value of its debts and its equity: $V = D + E$. Proposition I of the Modigliani-Miller (M-M) theorem [44] says that the market value of the firm is not dependent on its capital structure, if the following assumptions hold:

- There are no taxes
- The market is efficient (and consequently the bankruptcy costs are zero)
- Absence of asymmetric information

Therefore under these assumptions capital structure does not matter. On the contrary, when capital structure matters, at least one of the M-M assumptions is violated. Consequently the M-M assumptions can guide us in finding the determinants of an optimal capital structure.

The M-M theorem can be extended to an environment with taxes, where interest payments are a tax deductible item. The amount saved on taxes due to leverage is called the Tax Shield (TS) and can be expressed as $TS = T_C \cdot D$, where T_C is the corporate tax rate and D is the value of a perpetual debt. The tax shield is therefore increasing in the debt/equity ratio.

It was showed[13] that the second assumption is violated as well: financial distress and bankruptcy have direct and indirect costs, such as loss of costumers, suppliers, and employees due to uncertain future, need of immediate sale of assets at lower prices, expenses on experts, and so on. As higher leverage means higher interest payments and thus higher probability of not meeting them and falling into financial distress, the expected distress costs are increasing with higher leverage. The effect on the overall firm value is therefore the opposite as for the tax shield.

Asymmetric information—i.e. the violation of the third assumption—implies agency costs, when the conflict of interest between different groups of stakeholders causes suboptimal investment decisions.[14] The typical examples of agency costs are over-investment, under-investment, and cashing-out problem, all of them gaining in significance in states of (or close to) financial distress. The negative effects of agency costs are increasing in leverage, and therefore shifting the optimal indebtedness to lower values.

3.2 Absolute Priority Rule

Absolute Priority Rule (APR) is a concept that describes how the assets should be divided among stakeholders after the event of bankruptcy. The basic order of the APR is, that a junior creditor receives some fraction of the remaining assets only in the case when senior creditors are paid in full. Similarly, equity holders receive nothing, unless all the creditors (both secured and unsecured) get the whole amount of their claim. Furthermore, when a class of stakeholders have the same seniority, they all receive the same ratio of their principal.

A considerable amount of literature[15] has been published on the violations of the APR: while under Chapter 7 (United States Bankruptcy Code) liquidation absolute priority is generally enforced, in the case of Chapter 11 reorganizations[16] violation of APR is rather a rule than an exception. The reason is, that equity holders have the power to enforce APR deviation during workout negotiations due to the structure of Chapter 11 rules. The management can put the firm in Chapter 11 at a moment when it is in the best interest of equity holders. As there is an automatic stay on payouts to claimants under Chapter 11, a renegotiation could enhance the situation of both equity and debt holders. In addition, the reorganization plan needs to be accepted by the shareholders as well, and therefore they can prolong the bargaining process, and therefore increase the costs of default. This is clearly not in the interest of the senior claimants, and so they rather distribute some value to equity holders and avoid long negotiations. For further discussion of optimality of negotiations during bankruptcy procedures see [27].

[13] See [47], or [10].

[14] More on this see, for example [2].

[15] See, for example, [25, 26, 39, 43, 54].

[16] See [18, 55].

A large amount of empirical research have been done in the past two decades about the consequences of these absolute priority violations, and the result showed that APR deviations are beneficial ex ante. They decrease the severity of over-investment in assets requiring managers' special skills and under-investment in firm-specific human capital [5], might improve the timing of bankruptcy [48], hold back excessive risk taking [19] and help to resolve under-investment problem [56]. On the other hand, negative effects of absolute priority violation arise through the problem of moral hazard with respect to investment decisions [4].

3.3 Game Theory Analysis of Credit Contracts

As a typical company of our interest has complex capital structure with many parties of interest, it is reasonable to examine the problem of financing from the perspective of Game Theory. This section is therefore dedicated to this topic, and is particularly based on the work of Ziegler [57]. Our paper may be viewed as additonal, complementary, approach to the game theory analysis of the corporate bankruptcy provided by [28].

The method combines game theory and option pricing, so the maximized value of an option (note the parallel of options and credit contracts) can be calculated. The essence of the method is a three-step procedure:

1. The game between players is defined. The game tree is constructed.
2. The uncertain payoffs are valued using option pricing theory, where the parameters are the player's possible actions.
3. The game is solved using backward induction or subgame perfection.

The strengths of such a method are: taking into account the time value of money and the market price of risk, and separating the valuation problem from the analysis of strategic interaction.

3.3.1 Credit and Collateral

In financial contracting two forms of moral hazard occur: risk-shifting in the situation of hidden action, and observability problem in the situation of hidden information. In the following text these two basic problems are analyzed, whereas more complicated issues will be addressed in the upcoming parts of the section.

The Risk-Shifting Problem The origin of the risk-shifting problem is the borrower's incentive to influence the risk of the project, as he could increase his expected payoff on the expense of the lender. If he is able to change the risk of the project without the creditor's notice, we are talking about hidden action. The lender usually anticipates such behaviour, and requires higher interest rate that leads to adverse selection [50]. An alternative solution is to closely monitor the activities of the borrower, however this increases the costs of lending and therefore the interest rate. The best option would be a contract designed in a way that the borrower has no incentive for risk-shifting without the need of monitoring.

Ziegler [57] examined the situation when the borrower is able to set the riskiness of the project after the debt contract have been signed and the final payoff is observable to both parties with no cost. As it turned out, there exists an infinite number of contracts that preclude risk-shifting, however only contracts with proportional payout are renegotiation-proof (i.e. a situation, when a renegotiation is desirable for both the creditor and the debtor cannot occur). Renegotiation usually involves costs, and therefore both parties will have an incentive to agree on a contract that is not changed over its whole life. This means, that in the case of hidden action, only all-equity financing avoids risk-shifting.

The Observability Problem When the terminal value of the investment is not observable by both parties, a problem arises how the final transfer should be determined. In fact, it can be expected in many situations, that the borrower will have more accurate information about the terminal value, and therefore he can report distorted figures to minimize his payout to the lender.

According to Townsend's [51] costly state verification model—where the lender and the borrower agree in advance on situations when the verification should be taken—the optimal contract has the following properties (pure strategies allowed only):

- If verification does not take place, the payment to the lender is equal to some constant amount D.
- Verification should be taken when the terminal value is below some pre-defined threshold.

This contract is similar to a debt contract with fixed payment D and verification as a parallel to declaration of bankruptcy. Thus the observability problem can be addressed with constant promised payment in no-bankruptcy states. As risk-shifting can be solved only by proportional payment, there is no contract that could avoid both problems simultaneously.

Collateral is an asset, that can be—according to the credit contract—seized in the event of default to limit the lender's losses. A considerable amount of literature has been published on the role of collateral in providing motivation for the borrower to avoid default. For instance, in [3], the loan repayment decision is dependent entirely on the relative values of the collateral and the amount of outstanding debt, default occurring if the value of the collateral at maturity is below the amount due. An inverse relationship between agency costs and the amount of collateral available to borrowers has been shown by [6].

Chan and Kanatas [13] mentioned two types of collateral: it is an existing asset (for example the financed project) or it is an additional asset, normally not available to the lender. Ziegler's model examines the effects of the latter, and concludes that risk-shifting problem disappears only when the loan is fully collateralized, resulting riskless loan. However, collateral protects the lender in two ways: grants higher recovery after bankruptcy and reduces the borrowers incentives to risk-shifting behaviour.

3.3.2 Endogenous Bankruptcy and Capital Structure

In the previous section the credit was a finite maturity contract with a single payment to the lender at maturity. Although such approach is good to understand project financing, it is less useful to model corporate financing. In reality firms keep operating by issuing new debt to finance their new projects, or to repay the maturing debt and therefore keep the ongoing projects alive. Bankruptcy happens, when the entity is unable to meet its contractual payments. In fact equity holders can decide at any point in time whether they want the firm to make the agreed payments or default and trigger bankruptcy. Thereby bankruptcy is an endogenous decision made by equity holders, even if it might be initiated in principle by the creditor.

Ziegler [57] analyses endogenous bankruptcy building on the base of Leland's [34] infinite horizon model with the introduction of several modifications. First, interest on the loan is divided to two distinct types, a continuous effective payment and an increase in the face value of the loan. This division allows to investigate the role of these two components in finding market equilibrium. Second, endogenous bankruptcy is discussed as a principal-agent problem and the agency costs of the equity holder's socially suboptimal behaviour are quantified. Third, the effect of loan covenants and information asymmetry are considered. Fourth, the properties of optimal capital structure are studied, and finally, an incentive contract is developed that could influence equity holder's bankruptcy choice.

The Model A lender and a borrower signs the following contract: at initial time the lender transfers a loan of F_0,[17] and in exchange the borrower pays instantaneous interest of $\phi D(t)dt$, where $D(t) = D_0 e^{\kappa t}$ is the face value of the debt at time t and ϕ is the instantaneous interest rate to be effectively paid on the perpetual debt. Asset sales are prohibited, therefore net cash outflows on interest payments are financed by equity dilution. As κ is the rate of increase in the face value of debt (and therefore the rate of increase in interest payments as well), it is assumed, that $\kappa < r$, where r is the risk-free interest rate.[18] Sinking fund corresponds to the setting $\kappa < 0$.

If (and only if) the debtor defaults on his interest payments, the firm is liquidated with costs proportional to the asset value. The creditor therefore receives $(1-\omega)S_B$ in the event of default, where ω is the proportion lost due to liquidation and S_B is asset value at the time of bankruptcy.

The game has the following structure:

1. The amount of debt, D_0, and interest rates κ and ϕ are determined, the contract is signed. In exchange for its promised obligations the firm receives the fair value of the loan, F_0.

[17] F_0 denotes the fair value of the loan at time 0, as it will be described in more details later.

[18] Otherwise the present value of the interest payments would converge to infinity.

2. The firm makes its investment decision with the associated risk, represented by the volatility rate, σ. In the financing of additional (later) projects under-investment problem might occur.
3. Equity holders choose their default strategy S_B. In the event of bankruptcy ωS_B is lost, $(1 - \omega)S_B$ is received by debt holders, and nothing remains to the equity holders.

The management is assumed to fully represent the equity holder's interest, hence there is no conflict of interest between these two parties. Reference [57] assumes the asset value, S to follow the usual geometric Brownian motion, and estimates the firm, equity and debt value using the standard framework based on Merton.

In line with the principle of backward induction, the last stage of the game is examined at first. In this step the equity holders choose optimal asset level S_B for triggering bankruptcy. This level can be found using first-order condition, and is equal to

$$S_B = \frac{(1 - \theta)\phi D(t)}{r - \kappa + \sigma^2/2},$$

where θ is the corporate tax rate.

As it can be noted, this optimal level is linear in $\phi D(t)$, and is independent on current asset value S. Furthermore, higher asset risk (σ) implies lower optimal bankruptcy boundary.

The Principal-Agent Problem and Agency Costs The principal-agent problem stems from the fact that the debtor (agent) adopts a different bankruptcy barrier than it would be optimal from the creditor's (principal's) view.[19] The creditor would choose a default boundary either to zero (to make his claim riskless) or as high as possible (to receive the firm's assets when they have a high value). The socially optimal bankruptcy strategy turns out to be the one with the lowest possible level of bankruptcy triggering, i.e. $S_B = 0$. This comes from the positive cost of bankruptcy for any asset value higher than zero.

In order to construct an incentive contract that would lead to socially optimal bankruptcy the effectively paid interest on debt, ϕ has to be zero, since for any other value the equity holders would trigger bankruptcy at a positive asset level. However, setting $\phi = 0$ means that the claim is worthless, as no interest is paid out. In other words, because of the borrower's limited liability, socially optimal default level can not be reached.

Armed with the above results the agency costs arising from endogenous bankruptcy can be expressed. The agency cost represents the expected deadweight loss caused by the expected costs of bankruptcy. Intuitively, these costs are in direct relationship with the probability of bankruptcy (increasing in S_B and $\phi D(t)$), and with the proportional loss due to liquidation, ω.

[19] The optimal default levels from the debtor's and the creditor's points of view are derived in [57], pp. 48–49.

The Investment Decision - Under-investment and Risk-shifting Once we have investigated the equity holder's optimal bankruptcy decision S_B, we should examine their investment choices. Two main issues are studied in the following paragraphs: under-investment and risk-shifting. Myers [45] highlighted that firms may abandon profitable projects in the existence of debt by refusing recapitalization of the firm. The reason of doing so is, that although equity holders would bear the full costs of the project, debt holders also benefit from this investment as the debt becomes less risky.

Ziegler [57] analyses the under-investment problem with a model that represents new investment as a scale up of the existing operations by some factor $w > 0$. The investment requires therefore additional wS of funding and increases the value of the firm's assets to $(1 + w)S$. Since additional (equity funded) investment reduces expected bankruptcy costs and increases tax shield,[20] it always increases the overall firm value.

The model's calculated change in the value of the equity shows, that it is always lower than the costs of the investment, and therefore the overall return to equity holders is negative. Hence under-investment always arises. This problem can be addressed by renegotiation of the debt (reduction of D, ϕ, or κ) in order to ensure positive expected return on investment for the equity holders, or alternatively by sharing the costs of the new investment.

So far in the model of endogenous bankruptcy constant and known asset risk σ was considered, however in some cases this assumption might not hold. The question is, whether the agent has an incentive to increase the asset risk if the principal can not observe (and therefore control) his action. To answer this, Ziegler examined the partial derivative of the equity value with respect to σ^2. The result shows, that a leveraged firm has always incentives to increase asset risk. This has an implication for the optimal behaviour of the lender: he should focus on monitoring asset risk instead of asset value, as the risk is the relevant variable for the borrowers' bankruptcy decision.

Agency costs of risk-shifting can be expressed as a difference between the firm value at the social optimum less the firm value with the possibility of risk-shifting. Since firm value decreases with bankruptcy costs, it can be maximized by setting these costs to zero by approaching σ to zero. Agency cost is therefore equal to

$$C = \lim_{\sigma \to 0} W(S) - \lim_{\sigma \to \infty} W(S),$$

where, again firm value is W. As Ziegler [57] showed, the difference in the above limits is

$$C = \frac{\theta \phi D(t)}{r - \kappa},$$

i.e. to the value of the (safe) tax shields.

[20] Note that early bankruptcy means no tax deductibility in the future, and therefore it decreases the current value of the tax shield.

Effects of Loan Covenants It was shown in the previous sections, that under certain conditions, a "plain vanilla" debt contract[21] might imply deadweight loss that moves the resulting firm value below its socially optimal level. To mitigate these losses, loan covenants might be introduced. A loan covenant is a condition agreed at debt issue that has to be fulfilled by the debtor. Covenants can take many forms, regulating operating activity, asset sale, cash payout and others.[22] Here, so-called safety covenants are analyzed which give the bondholder the right to force bankruptcy if certain conditions are met. More specifically, suppose a covenant that forces the firm into bankruptcy, if its asset value falls below some specified level $\overline{S_B}$. Reaching this level means transfer the ownership of the assets to the lender. As it turned out, the risk-shifting incentive depends on the level of this barrier: for low levels risk-shifting incentive is still present, however for higher values the situation changes and the debtor will have an incentive to decrease the risk of the investment. The breakpoint is naturally higher than S_B, the endogenous bankruptcy barrier set by the equity holders only.[23] Concluding the effects of such loan covenant, we should remark that they protect the lenders in two ways: First, they reduce losses of the creditors by setting the default barrier higher, and Second, they mitigate or even eliminate equity holder's risk-shifting incentives. Hence, setting a safety covenant with an agreed level has similar effects as using collateral.

3.3.3 The Financing Decision

Using the results derived, we can investigate the way a firm should be financed. We will analyse—under endogenous bankruptcy—the optimal capital structure of a firm, and the effects of the way how the interest is divided between the interest effectively paid and growth rate in the face value of debt.

Optimal Capital Structure Assume that the asset risk is known to the lender and risk-shifting is not possible, or alternatively, it is possible only within certain bounds. In the latter case the lender would anticipate the borrower's risk-shifting behaviour, and therefore he will use the maximal available volatility value in his loan pricing calculations, $\bar{\sigma}$. We assume that the face value of the loan cannot be changed after the initial agreement, and that the borrower takes the offered interest rates κ and ϕ as given when selecting the initial face value of debt, D_0.

The financing decision is made with respect to the equity holders' effort to maximize the value of their holdings after the initial investment, I. Ziegler's calculations show, that there exists an interior maximum of the net equity value (that is the difference between the value of equity after the debt is taken and the equity holders' initial investment) in terms of optimal capital structure. As the rate of effective

[21] Here "plain vanilla" refers to the absence of additional clauses defining loan covenants.

[22] A comprehensive analysis of covenants and their effect on debt pricing can be found in the work of Reisel [49].

[23] For the mathematical derivation of this statement see [57], pp. 58–59.

interest payments, ϕ rises—and consequently so does the cost of the debt service—the optimal face value of debt decreases. Similarly a higher growth rate in the face value of debt, κ, means lower optimal face value of debt. It also turns out, that changes in ϕ are perfectly offset by the endogenously chosen face value of the debt, and so the continuously paid coupon remains the same. Thus ϕ affects the nominal leverage (D_0/S_0), however it does not affect the leverage in market terms (F_0/S_0).

Interest Payments Versus Increase in the Face Value of Debt A natural question is, how the debt service should be divided between the interest payments ϕ, and the growth rate of face value of the debt κ. As the optimal leverage in market terms is not affected by ϕ, the borrower is indifferent to the interest rate effectively paid. On contrary, the rate κ does affect the optimal capital structure and the net equity value: with increasing κ the optimal leverage ratio and the net equity value decreases. Consequently equity holders prefer to pay higher effective interest instead of higher growth in the face value of debt.

Expected Life of Companies As the optimal capital structure and the conditions of the loan are given, it is possible to express the mean time of default. Using the analysis of Ingersoll [24], we know that the mean time of passing the origin for a standard geometric Brownian motion $dx = \mu dt + \sigma dW_t$ with initial value x_0 is given by

$$\bar{\tau} = \frac{x_0}{\mu}$$

With the help of this formula—after some computations—the mean time of default under endogenous bankruptcy can be revealed.[24] This value turns out to be independent on the parameter ϕ, in line with the finding that the borrower offsets the changes in the effective payout rate by changing the face value of debt. Again, the important parameter is κ, that influences mean time to bankruptcy.

An Incentive Contract It is worth to consider whether the lender can set the contract parameters ϕ and κ in a way that influences the borrower's bankruptcy strategy S_B. As it is in the lender's interest to have a higher default barrier, we will examine the possibilities of an incentive contract that induces the borrower to declare bankruptcy at a higher asset value. Early bankruptcy is interesting for the borrower for several reasons. First, the lender might be himself an agent and so he might have restrictions on the maximum risk he can take. Second, early liquidation may increase beliefs about the lender's solvency and therefore avoid some problems such as bank runs. Third, it enables the lender to save on monitoring costs as he can use early information provided by default on interest payments.

Since changes in the effective interest rate are perfectly offset by changes in D_0, ϕ does not affect the borrower's behavior. On the other hand the rate of growth in the face value of debt, κ does influence the borrower's optimal bankruptcy strategy.

[24] See [57], pp. 67–68.

As a lower κ means faster debt repayment (through higher face value or equivalently higher ϕ), the resulting optimal bankruptcy triggering level is higher.

4 A Proposed Structural Model

Section 3 gave an insight to the design of credit contracts, and showed the usability of game theory in pricing of corporate assets and predictions of rational actions taken by the parties concerned. Here, we extend the available literature of asset pricing models introduced in Sect. 2, and build up a framework with stochastic interest rate. This framework than serves as a valuation method for a similar game theory analysis as was introduced in Sect. 3.3. The starting-point of this work is the Goldstein et al. [21] EBIT-based model, that will be extended by the relaxation of the constant (or deterministic) interest rate requirement.

4.1 Assumptions

First of all we take the following assumptions:

(i) The management fully represents the equity holders' interest.
(ii) The APR is never violated.
(iii) Asset sales are prohibited, interest payments are financed by earnings and equity dilution.
(iv) When the earnings are above the paid interest, the difference is paid out as dividend.
(v) Paid interest is a tax deductible item, however no tax carry-back or carry-forward exists.[25]
(vi) There is a sufficiently large number of investors, and only a limited amount of projects.

Assumptions (iii), (iv), and (v) imply the unimportance of the historical cash flow in the asset pricing. The current values of the two memoryless processes—the risk-free interest rate and the EBIT—are the only two stochastic variables that affect the debt, equity and firm value. Assumption (vi) has the consequence that the provided loan is always fairly priced, since the financial institutions perfectly compete with each other. Next to these initial assumptions we will use further suppositions in the subsequent sections, particularly during the description of the stochastic evolution of the variables: the risk-free interest follows an Ornstein-Uhlenbeck process, the Earnings Before Interest and Taxes (EBIT) is supposed to follow a GBM, and so on.

[25] As Nejadmalayeri and Singh [46] showed, the US tax code's loss carry provisions affect the equity holders' bankruptcy decision.

4.2 Risk-Free Interest Rate

Most of the models assume constant risk-free interest rate in order to simplify the calculation. However, in reality this interest rate does change in time, reflecting the situation of the overall economy. Modelling the interest rate stochastically allows us to include the possibility of a macro-level change and catch the correlation between the overall market and the modelled asset. Using this correlation the model could be extended to a risk averse measure, where higher return is expected just for the market risk—the one that can not be diversified (in line with modern portfolio theory, see [38]).

The risk-free interest rate $r(t)$ follows an Ornstein-Uhlenbeck process suggested by Vasicek [52], and used for example in the Longstaff and Schwartz [36] approach:

$$dr = \alpha(\gamma - r)dt + \sigma_r dW_t \tag{17}$$

where $\alpha > 0$ indicates the force pulling the interest rate back to its long-term mean γ at speed $\alpha(\gamma - r)$ per unit of time. The stochastic element is a standard Wiener process W_t times the volatility σ_r.

The expected value and variance at time s given $r(t)$ are

$$\begin{aligned} E_t[r(s)] &= \gamma + (r(t) - \gamma)e^{-\alpha(s-t)}, \quad t \le s \\ Var_t[r(s)] &= \frac{\sigma_r^2}{2\alpha}(1 - e^{-2\alpha(s-t)}), \quad t \le s \end{aligned}$$

respectively. The distribution of $r(s)$ given $r(t)$, $t \le s$ can be written as

$$r(s) = r(t)e^{-\alpha(s-t)} + \gamma(1 - e^{-\alpha(s-t)}) + \frac{\sigma_r}{\sqrt{2\alpha}} W_t(e^{2\alpha(s-t)} - 1)e^{-\alpha(s-t)}$$

Having the assumption of risk-neutral measure (i.e. the yield to maturity is not dependent on the maturity date and thus there is no risk premium), the value of $1 received at time $s \ge t$ has the value of

$$P(t, s) = E_t\left[\exp\left\{-\int_t^s r(\tau)d\tau\right\}\right] \tag{18}$$

received at t.

4.3 Earnings Before Interest and Taxes

Traditional models—building on the basis of Merton's [42] framework, including those introduced in Sect. 2—take unlevered equity as primitive variable with log-normal dynamics. However, for some models it seems to be more straightforward to use earnings instead of unlevered equity. Mella-Barral and Perraudin [40] consider a firm that produces output and sells it on the market, where the price of the sold product follows a geometric Brownian motion. Mello and Parsons [41] use a similar framework with a mining company and stochastic commodity price movements. Graham [22] models EBIT flow as a pseudo-random walk with drift, Goldstein et al. [21] and Broadie et al. [12] use geometric Brownian motion for the evolution of EBIT.

To see the advantages of such approach, we should review some of the main shortcomings of the traditional framework. **First**, unlevered equity ceases to exist as a traded asset when debt is issued. This problem is one of the motivating factors behind several subsequent frameworks [17, 31, 32]. **Second**, they treat tax payments in a different fashion as they deal with cash flows to debt and equity holders. In fact, they count tax benefit as capital inflow instead of using it for reduction of outflows. This implicitly assumes that it is always possible to deduce fully the interest costs from the tax payments, however, this is not the case when the cost of debt service is higher than the current EBIT. Another problem with the tax benefit approach is, that it implies higher firm value through higher tax shield as the tax rate increases. **Third**, as Goldstein et al. [21] noted, these models may significantly overestimate the risk-neutral drift, consequently underestimate the probability of bankruptcy and so the optimal leverage ratio.

Our model assumes an EBIT process with log-normal dynamics, and therefore is able to address the mentioned issues. The evolution of the firm's instantaneous EBIT, δ_t is modeled using geometric Brownian motion with risk-neutral measure $\mathbb{Q}$, similarly as Broadie et al. [12]:

$$\frac{d\delta_t}{\delta_t} = \mu dt + \sigma dX_t(\mathbb{Q}), \tag{19}$$

where

$$X_t = \rho W_t + \sqrt{(1-\rho^2)} Z_t.$$

W_t is the same process as in (17), Z_t is a standard Wiener process and ρ is the correlation coefficient between the risk-free interest rate and EBIT.

If the δ_t is known at $t = 0$, the differential equation (19) has the solution

$$\delta_t = \delta_0 \cdot \exp\left\{\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma X_t\right\} \tag{20}$$

Assuming no taxes and zero leverage, the value of the firm is the sum of discounted earnings. Using the notation V_t^0 for unlevered equity value at time t, we have

$$V_t^0 = \int_t^\infty \delta_t \cdot \exp\left\{\left(\mu - \frac{\sigma^2}{2}\right)(s-t) + \sigma X_s - \int_t^s r(\tau) d\tau\right\} ds,$$

in line with (18).

4.4 Debt

The debt issuance and repayment is similar as in Ziegler's [57] model with endogenous bankruptcy, although several modifications are implemented. Most importantly, as the risk-free interest rate is considered to be stochastic, the interest payments are stochastic as well. Second, Ziegler considered a debt service divided between effective interest payments and growth in Face Value of debt (FV). As he proved that changes in effective interest rate are compensated by changes in face value of debt, its scalability will be left out from our model.

The debt is therefore set up it the following way:

1. The rate of growth in face value of debt, κ is chosen
2. The borrower (i.e. the firm) chooses the initial face value of debt, FV_0
3. The lender calculates the fair value of this debt, given the face value and κ, and provides a transfer to the borrower equal to this fair value.

After receiving the funds, the borrower starts to serve the interest payments. The FV at any point in time is given as:

$$FV_t = FV_0 \cdot e^{\kappa t}$$

The interest is continuously paid out at a rate $c_t = FV_t \cdot r(t)$ (coupon rate) with infinite horizon. We assume $\kappa < \gamma$, similarly as Ziegler, otherwise the discounted FV, and consequently the interest payments would growth to infinity.

The economic intuition behind this model is a floating coupon perpetual bond issue, where this corporate bond is (usually) sold below par. In order to catch constructions as a sinking fund, or alternatively a growth in debt principal, the parameter κ is introduced as well.

4.5 Default

The event of default corresponds to the situation, when the firm does not meet its obligation on interest payments. We assume, that creditors take over the firm immediately after its default and suffer the associated losses. Absolute priority rule is enforced, i.e. after bankruptcy equity holders receive nothing.

As the state variable is the instantaneous EBIT, it is convenient to define the recovery value as a multiple of the EBIT at the moment of default. Since a firm effectively becomes unlevered after bankruptcy (as its debt holders become the new equity holders), and we calculate the unlevered value during the iterations, this multiplier can be easily transformed to Loss Given Default (LGD)—a ratio that expresses the asset value lost due to bankruptcy.

4.5.1 Default Barrier

It is sensible to define the Default Barrier (DB) on the state (primitive) variable, since all the other values can be written as a function of this state variable. As we have an EBIT based model, DB will be defined on earnings. When the primitive variable is firm (or unlevered equity) value, DB is usually a function of the face value of debt. A straightforward modification for our model is to make the DB linearly dependent on the instantaneous coupon rate, c_t (as we show in Sect. 4.9, this setup proved to be consistent with the overall model).

Such modification would imply a lower barrier in recession (low risk-free rate), and thus work counter-cyclically. There are several facts that support this design: in recession the number of bankruptcies increases (see, for example [1]), thus banks experience losses in connection with other loans and might prefer immediate payments instead of triggering bankruptcy that yields uncertain income later. Furthermore as Altman et al. [1] also showed, the recovery rate is significantly lower in recession. Exactly the opposite holds for economic boom and high interest rates, therefore higher default barrier is reasonable.

4.5.2 The Bankruptcy Decision

The entity that does the bankruptcy decision is dependent on the transparency of the firm, on the credit contract, and possibly on other factors. When the state variable is not publicly observable, the firm's management is the only one who can trigger bankruptcy. On the contrary, when the state variable is observable, bankruptcy decision can be declared in the credit contract, and therefore support more favourable debt financing. This is in fact a safety covenant for the creditors, that ensures them the right to force bankruptcy if the firm performs poorly (that is crosses the DB).

4.6 Method and Calculations

Due to the high complexity of the model we use Monte Carlo simulations to uncover the model's sensitivity on its parameters (see Table 1 for parameter base values). The calculated results are used as payoff valuation for game trees analyzed in Sect. 4.7.

4.6.1 The Effects of Debt Face Value

The Face Value of debt (FV) is the most basic parameter of a corporate loan: it is the figure that appears on the firm's balance sheet and in other reports and statistics. It is also the exclusive right of the borrower to specify the loan's FV directly or through the amount of borrowed funds. The main questions addressed in the following lines are, whether it pays off to issue debt at all, whether there is a maximal firm value and

Table 1 Notation

Symbol	Explanation	Base value
Interest rate		
$r(t)$	Risk-free interest rate	$r(0) = \gamma$
γ	Long-term mean of risk-free interest rate	3%
α	Speed of expected risk-free interest rate convergence to γ	0.25
σ_r	The volatility of risk-free interest rate	0.5%
$P(t, s)$	The price of a $1 face value riskless zero-coupon bond at time t, maturing at time s	
Firm		
δ_t	EBIT	$\delta_0 = 100$
μ	Drift of EBIT under $\mathbb{Q}$	0.01
σ	Volatility of EBIT	20%
ρ	Correlation coefficient between $r(t)$ and δ_t	0.2
V^0	Firm value with no leverage and the assumption of zero taxes	
T_C	Corporate tax rate	35%
Debt		
FV_t	Face value of debt	
κ	Growth rate of the face value of debt FV_t	1%
$D(\delta_t)$	Debt value	
c_t	Coupon rate, equals to $FV_t \cdot r(t)$	
Default		
DB_t	Default Barrier	
τ	Time of default	
RR	Recovery rate defined as a multiple of yearly EBIT	10×

if so, what level of FV corresponds to this maximum, and how this optimal value is dependent on the DB.[26]

Figure 1 illustrates the dependence of debt, equity and firm values on credit contracts with different face values. As it is visible, when the leverage is low, firm value can be enhanced if a debt with higher face value is issued due to increasing tax shield. At a certain point the rising bankruptcy costs exceed further tax savings, indicating an optimal face value of debt that maximizes firm value. With a low DB[27] equal to 0.3, for example the firm value can reach 35 times the yearly EBIT if a debt is issued with face value between 20 and 30 yearly earnings. This means an optimal debt ratio of circa 60–80%. As the DB rises, this optimal ratio declines due to higher

[26] At this point we do not concentrate on the problem how the DB is chosen; that issue will be covered in Sect. 4.7.

[27] Recall that a default barrier of 0.3 means triggering default when the instantaneous earnings are at 30% of the coupon rate.

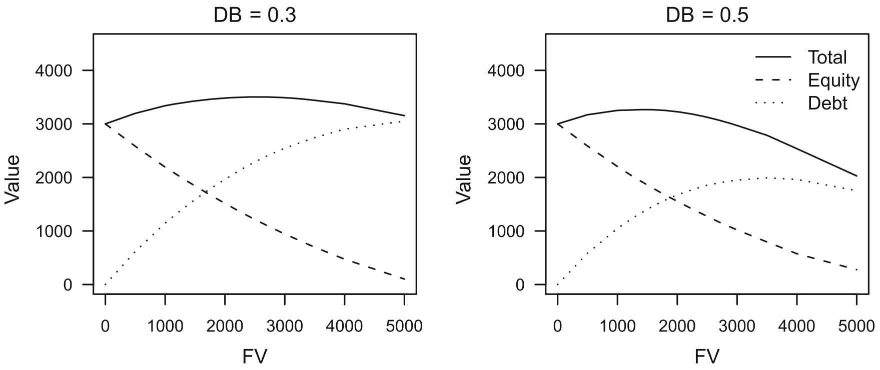


Fig. 1 Debt, equity and total value with different face values of debt

Probability of Default (PD): with $DB = 0.7$ the maximal firm value declines below 3200 (i.e. 32 times the yearly EBIT) with debt ratio of 30% only. The effects of changes in the DB are described in details in Sect. 4.6.2.

By observing the debt values it is apparent that at a certain FV the debt value reaches its maximum: this is the highest possible amount of money that could be reached with sole debt financing. The plotted equity values are not relevant as the equity holders are compensated for their decrease in equity value by receiving the funds obtained from the loan. Therefore the equity holders seek a loan agreement that ex-post maximizes firm value.[28]

4.6.2 The Effects of Default Barrier Level

Next, we should explore how the output variables react on different levels of default barriers. To do so, we have plotted our basic calculation,[29] where no extreme values distort the picture. Figure 2 shows how the level of default barrier affects the equity, debt and overall firm value.

[28] This holds only at the moment when the contract is signed. Later on both the debt and equity holders profit from an increase in the firm value.

[29] That is the one with parameters set to their base levels.

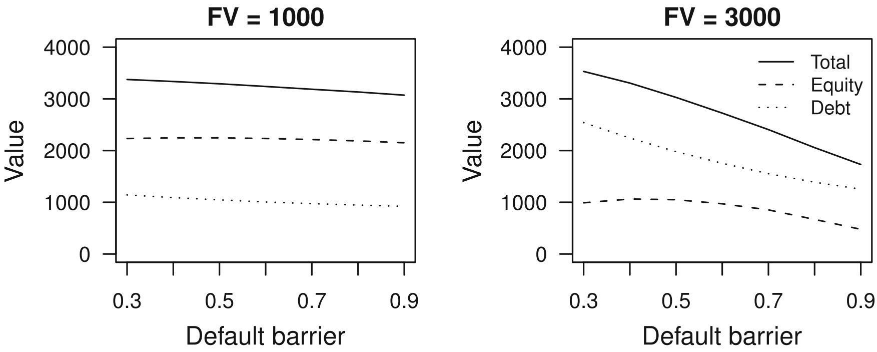


Fig. 2 Debt, equity and total value dependence on the DB with FV 1000 and 2750

The overall firm value has the most unequivocal trend: it is declining as the barrier rises: the FV affects only the slope, not the tendency. Intuitively, setting the DB lower implies drop in the number of bankruptcies, later occurrence of the expected bankruptcy, and shrink of the LGD in absolute terms. Recall that the expected costs of bankruptcy equal to the product of these three factors: PD, LGD and the discount.

The value of debt is rising with lower DB level. Again, this is intuitive, since default occurs later, therefore more money flows to creditors through equity dilution. If we examine the curves of the debt value on Fig. 2, a convergence in this value can be observed, as the DB rises. Because the initial EBIT is set to 100 and the base value of the RR multiple is 10, the debt value needs to be 1000 for sufficiently high DB that triggers default immediately. Consequently this needs to be the level where debt value converges to.

The third curve—the one that demonstrates the equity value sensitivity on shifts in the DB—is somewhat different: it has a "quadratic" shape with a maximum around 0.5. This means that, from the equity holders' point of view, there exists an optimal non-zero default decision. This result is highly important for our game theory analysis in Sect. 4.7, where we examine the rational behaviour of the involved parties. This conclusion, as well as the results related to the firm and debt values, is in line with Ziegler's [57] findings derived using closed form calculations in constant interest rate environment.

4.7 *Agency Costs*

4.7.1 Observable Actions

With observable actions, the creditor is able to control the parameters that affect the probability distribution of the EBIT flow, most importantly σ, which is determined by the riskiness of the firm's projects. This situation significantly simplifies the arrangement of the credit contract, since the lender does not need to study the set of possible actions that might be done by the debtor. In other words, the probability distribution of the payoffs is given, and therefore risk-shifting is not possible.[30]

Observable State Variable The simplest situation is, when the firm is completely transparent, and therefore the creditor can observe the management's actions and also the state of the firm. In this case a debt contract can be signed with such covenants that enforce both an agreed volatility and defines a default barrier at which bankruptcy will be triggered.

In this case such a combination of debt face value and default barrier will be chosen that maximizes firm value. This leads to a highly leveraged firm (to maximize the value of tax shield), and to low default barrier (to minimize the bankruptcy costs). Note, that it might be not always possible to specify an arbitrarily low DB: when the EBIT decreases so drastically, that the equity becomes worthless, it is not possible to finance the interest payments trough equity dilution. In a stock company the shareholders cannot be forced to transfer additional funds to the distressed firm. In contrast, when the considered firm is owned by a parent company, the interest payments can be guaranteed by the mother.

Not Observable State Variable Similarly as in the previous case, actions are observable, and therefore risk shifting is not possible. However, as the state variable is not followed by the creditor, a bankruptcy barrier as safety covenant can not be included in the credit contract, because it would be impossible to enforce it. Consequently the debtor will choose the default barrier in a way that maximizes its equity holders' value under the given circumstances. This decision is the bottom level of the game tree, and therefore it determines the expected payoffs under certain credit contract parameters. Table 2 shows an equity value matrix for several debt face values calculated using the base parameter setting.[31] As it can seen, the equity holders will choose to default on interest payments when the EBIT will be between 40 and 50% of the coupon rate (bold values in Table 2).

As the lender anticipates the borrower's behaviour in the bankruptcy triggering decision, he prices the loan according to this action. We have discussed in Sect. 4.6.1, that the equity holders want to maximize the overall firm value, and so they will choose FV that implies this highest possible value. Table 3 gives the valuation of this step in the game: the creditor offers loans priced according to the equity holders's default decision, therefore the equity holders' can choose total firm value only within the column specified by the planned (by shareholders) respectively assumed (by bondholders) DB. In this case the optimal face value of debt is 2000 for $DB = 0.4$ and 1500 for $DB = 0.5$. The corresponding firm values are 3400 and 3300 respectively.[32]

[30] More about risk shifting in the next section, where—in contrast with the present situation—it is possible.

[31] See Table 1.

[32] All these values are rounded: as we want to illustrate the decision process, the accurate numbers are not important. In real the DB is one number (between the mentioned 0.4 and 0.5) not an interval,

Table 2 Equity values - basic parameters

	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	3037	3037	3037	3037	3037	3037	3037
500	2618	**2623**	**2623**	2619	2614	2604	2593
1000	2233	**2246**	**2246**	2234	2213	2187	2151
1500	1881	**1904**	**1903**	1881	1848	1792	1727
2000	1558	**1595**	**1595**	1562	1504	1404	1281
2500	1264	**1319**	**1316**	1257	1162	1037	865
3000	990	**1063**	**1050**	971	853	669	477
3500	742	**837**	**813**	725	563	354	121
4000	510	**623**	**610**	483	304	73	0

Default barrier on the X-axis and debt face value on the Y-axis

Table 3 Total firm values - basic parameters

	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	3037	3037	3037	3037	3037	3037	3037
500	3234	3222	3207	3191	3175	3156	3137
1000	3375	3336	3292	3240	3186	3133	3072
1500	3471	3393	**3306**	3211	3118	3010	2897
2000	3525	**3404**	3270	3122	2966	2778	2582
2500	3543	3375	3173	2952	2714	2463	2179
3000	3531	3306	3029	2725	2406	2057	1731
3500	3499	3203	2839	2454	2049	1632	1203
4000	3429	3045	2613	2123	1650	1147	1000

Default barrier on the X-axis and debt face value on the Y-axis

The resulting total value, equal to 33–34 yearly EBITs is significantly higher than the unlevered value with 30 EBITs only. On the other hand, the maximally possible 3550 is not reached due to agency costs caused by asymmetric information.

Paradoxically, the equity holders' ex post effort to increase the value of their claim decreases the total firm value (and so their total payoff) ex ante. This problem can be solved if they manage to ensure the lender, that they will default on their payments when the EBIT truly crosses the DB. Such contract requires monitoring with some associated costs, however if these costs are below the agency costs then monitoring should be introduced.

(Footnote 32 continued)
and the FV that corresponds to the maximal firm value given this DB is determined unambiguously as well.

4.7.2 Hidden Actions

When the management's actions are not observable, the debtor is able to modify the parameters driving the EBIT flow, and so to change the expected payoffs of the involved parties. More specifically, he is able to shift the risk to the creditor, and consequently to enhance the value of his claim on the creditor's costs. Such behaviour is called risk-shifting or, in a wider sense, moral hazard.

To demonstrate this problem, recall Sect. 2.1, where we described how Merton [42] proved that the value of equity in a leveraged firm can be expressed as European call option, and (using put-call parity) the value of debt is equal to a riskless bond with appropriate parameters less the value of a European put option. When the volatility of the asset's value rises, both options become more valuable, and therefore the equity value rises while the debt value declines. This model is valid only when there is no default prior debt maturity (and other assumptions made by Merton hold), however it illustrates the principle of risk-shifting.

To find out whether risk-shifting appears in our model, and if so, what are its consequences, we have run simulations with several different EBIT volatility parameters. For the details of the simulation see [15]. With higher σ values we observed the following (see Fig. 3):

Equity value was rising, with steeper slopes for lower DB settings. In consequence the equity holders try to increase the EBIT volatility as much as they can, however they have a lower incentive to do so when the DB is higher. This means that if there are some additional costs of higher volatility paid by the equity holders,[33] than they will not set the volatility to such high levels as they would so with lower DB.

Debt value was declining, however this decline was moderate for high DB settings. There are two reasons that support lower losses in debt value: First, and most importantly, default occurs at higher firm value, and therefore the firm has higher residual value after the bankruptcy that is transferred to the creditor. Second, default occurs earlier, therefore the asset value received has a smaller discount.

Probability of default rose.

Total firm value was decreasing due to increased PD.

Default barrier chosen by the equity holders was decreasing: their option on the firm's assets become more valuable with the increased volatility.

All of these observations are in line with the conclusions of Ziegler [57], who based his analysis on game theory and gave closed-form results for his model with constant risk-free interest rate. Next we examine how the observability of the instantaneous EBIT affects the credit contract's design and the behaviour of the involved parties.

Observable State Variable If the state variable is observable, it is feasible to mitigate the equity holders' risk-shifting incentive by setting a sufficiently high DB as a safety covenant. For a better understanding of the mechanism of this safety covenant we extend the Mertonian parallel of the equity value and a European call option.

[33] This could be lower expected EBIT growth, or some risk of being exposed, for example.

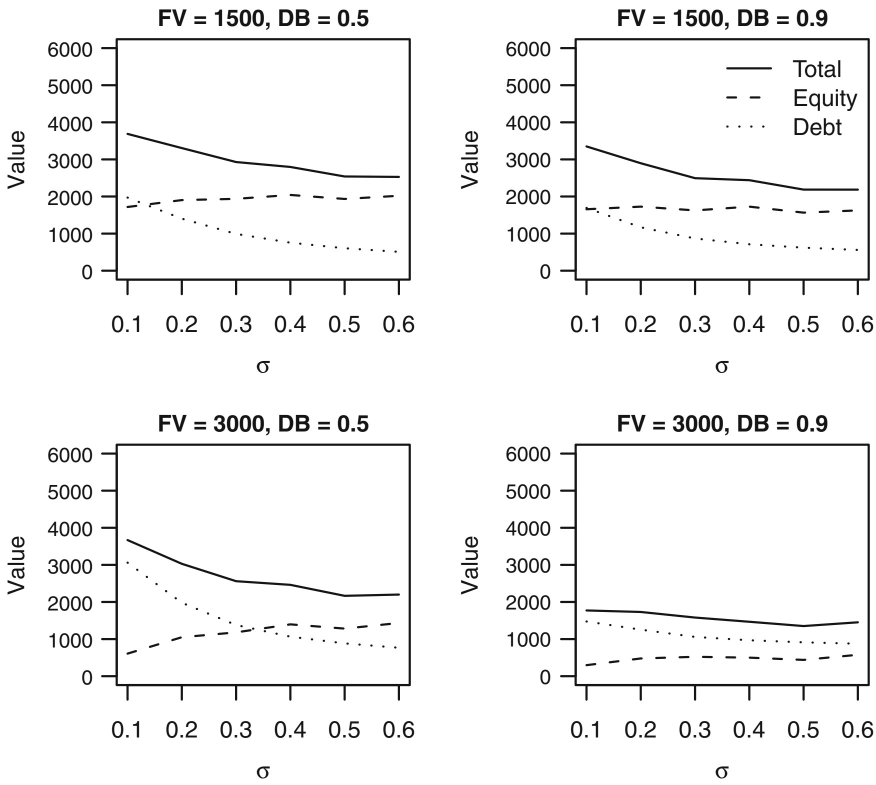


Fig. 3 Firm value dependence on σ

After the introduction of an exogenous default barrier the European call option is replaced by a down-and-out call barrier option.

Such an option has a similar price as a plain vanilla option if the DB is far below the spot price, and the volatility is not extremely high. However, as the spot price approaches the barrier, the option values begin to significantly differ. Figure 4 shows[34] the prices of down-and-out barrier and plain vanilla call options as a function of the volatility, assuming a strike price 1000, barrier 900, constant risk-free interest 3% and time to maturity 1 year. As we can see, the equity holders' incentive to increase the volatility is mitigated when the firm value approaches the DB.

Our model shows a similar behaviour: when the DB is high (80–90% of the coupon rate), the equity value is not increasing significantly with higher volatility. A high DB can be used therefore as a safety covenant in order to avoid risk-shifting. This implies a loan with low FV (about 5 yearly EBITs in our basic setting; recall Fig. 1), and consequently results a total firm value of only circa 3150 (31.5 yearly EBITs). Comparing this number with the theoretical maximum of a fully transparent firm

[34] Source: author's calculations using Financial Derivatives Toolbox.

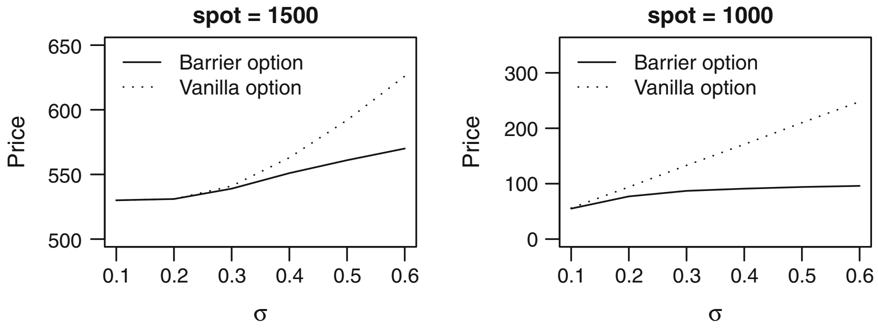


Fig. 4 Barrier option price dependency on volatility, barrier 90% of strike

(3550), the losses caused by risk-shifting are equal to the firm's four yearly earnings. Similarly as in the case of not observable state variable, it might pay off to introduce monitoring on the management's actions, and therefore to avoid risk-shifting.

Not Observable State Variable If the state variable is not observable, equity holders will increase the EBIT volatility and default on interest payments later. Since the creditor anticipates such behaviour, he prices the loan with respect to higher expected volatility. Consequently the resulting firm value (as it is depicted in Fig. 3) is lower than the value of the unlevered firm. The shareholders' ex-post behaviour therefore disables debt financing, and hence making the possible tax benefits unavailable.

4.8 Initial Interest Rate Level

An important advantage of the introduced mean-reverting interest rate environment is, that it can deal with a risk-free interest rate that is not on its long-term average (γ). In such case the interest rate is expected to return to γ, however, this takes some (random) time. In models with constant interest rate it is not possible to cover this situation. With a stochastic interest rate model though, it is just a question of different initial value $r(0)$ in the SDE (17). Furthermore, the effects of exogenous changes in this initial level can be examined. These exogenous changes in the risk-free interest rate correspond to the decisions of the central bank, and therefore we are able to predict the effects of the monetary policy on microeconomical level.

To see the effects of changes in the initial interest rate, we have run calculations with $r(0) = 1\%$, $r(0) = 3\%$, and $r(0) = 5\%$. Figure 5 demonstrates the obtained results for two different FVs. The tick lines show the total firm value dependence on the DB for three different initial interest rate levels. The gap between these lines represent the loss—ceteris paribus—when the interest rate suddenly increases to the next examined level. This drop in firm value is caused by two factors: higher

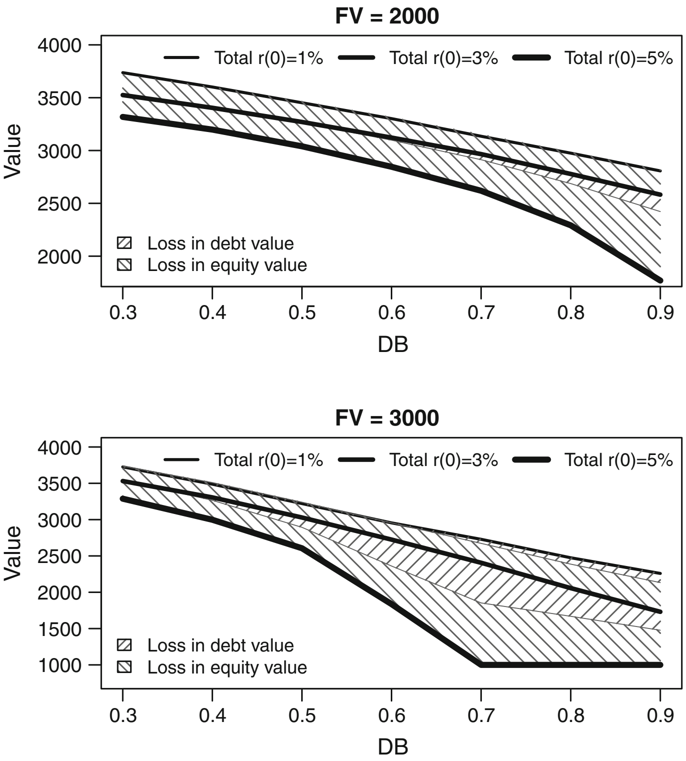


Fig. 5 Firm value dependence on initial interest rate

discount for all future earnings and increased PD due to higher interest payments.[35] The mentioned gap is a sum of declines in equity and debt value, and therefore we can divide this area to distinguish the losses of the two involved parties.

A larger fraction of the firm losses is booked by the equity holders (recall Fig. 5). Their claim is depreciated by the factors that affect the firm value (i.e. higher discount of future income and increased PD), and also by one additional: higher interest paid out to debt holders.

We can see that the debt value is insensitive to changes in initial interest rate, when the probability of early default is close to zero due to low FV and DB. Our conclusion is, that increased coupon payments perfectly offset higher discount on

[35] Higher interest payments imply higher DB in absolute terms. The DB of the x axis on Fig. 5 is a ratio of the instantaneous interest payments.

future cash flows.[36] Consequently the only factor that decreases the bond's value is the increased default probability and its earlier expected occurrence.

Note, that this section explains how the central bank's interventions work. In economical downturn the monetary policy can support the companies by targeting a lower short-term rate. This increases the value of both traded and non-traded assets, reduces the number of defaults, and supports debt financing through the decrease of interest paid on the outstanding principal. The latter is favored by two factors: the risk-free interest is low, and the risk-premium drops due to lower PD. On the contrary, an overheated economy can be cooled down with higher risk-free interest.

4.9 Comparison of Stochastic and Deterministic Default Barrier

Stochastic risk-free interest rate and DB are the two features of our model that distinguish it from other EBIT-based works [12, 21]. The contribution of a stochastic interest rate is intuitive: a constant or deterministic risk-free rate is hardly acceptable. Its usefulness was presented also in Sect. 4.8, where our model have easily dealt with different initial interest rate levels and it was able to predict the implications of macro-level shocks. The benefits of a stochastic DB were however not proved. In the description of the DB for our model (see Sect. 4.5.1) we mentioned why banks might prefer a DB that is dependent on the interest rate. We saw however, that it is not the bank who sets the default triggering level: it is the debtor or it is specified in the debt contract, that is designed by both parties.

In order to examine whether it is correct to base our model on stochastic DB we simulated two firms with identical parameters[37] but different DB settings: one stochastic, driven by the instantaneous risk-free interest rate, and one deterministic DB, dependent only on FV_t.

The default triggering levels were therefore set to $FV_t \cdot r(t) \cdot DB$ in the stochastic case and to $FV_t \cdot \gamma \cdot DB$ in the deterministic case, where $DB > 0$ is the same variable in both cases. Figure 6 visualizes the comparison of results obtained by stochastic and deterministic DB setting. For the first sight it is apparent that the total firm value is higher when the DB is defined as a deterministic function. The reason is that a deterministic DB in fact softens the default triggering bound, and hence increases the firm value. The problem is however, that when the primitive variable is

[36] For $\kappa = 0$ this is intuitive: the defaultable corporate bond can be represented as a risk-free bond with the same parameters minus the expected losses caused by default. Since the price of a riskless bond that pays continuous interest is always equal to its face value, it is not dependent on the current interest rate.

[37] These parameters were the same as in the basic setting, with the exception of lower recovery rate (5 yearly EBITs), and higher correlation between the EBIT and interest rate processes ($\rho = 0.5$). These modifications were made in order to make the results more sensible on the selection of the DB. Furthermore the number of iterations was doubled to increase the significance of small deviations between the two settings.

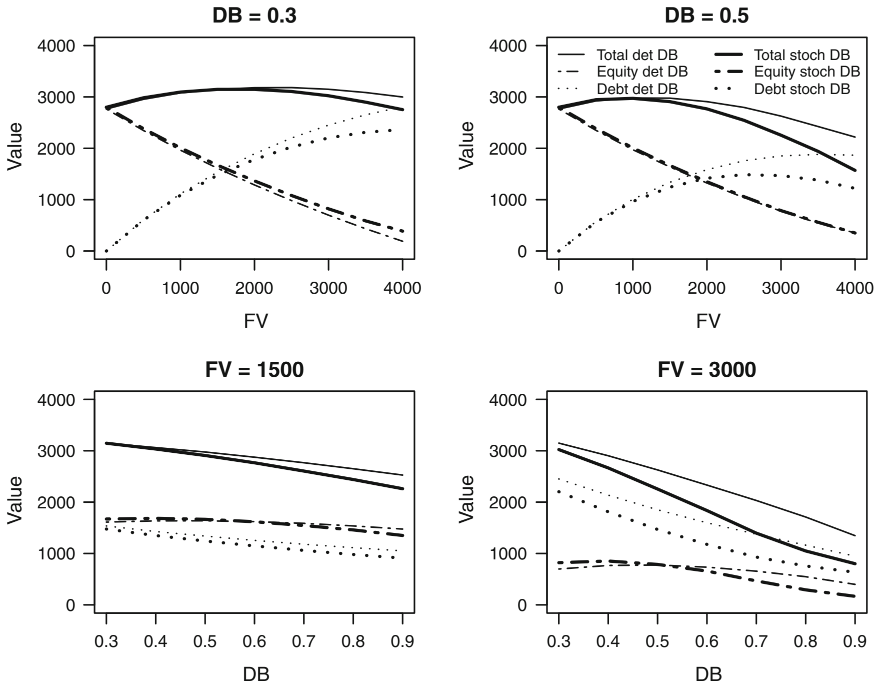


Fig. 6 Stochastic versus deterministic DB

not observable,[38] default is triggered by the equity holders in a way to maximize the value of their claim. Recall Fig. 6: a stochastic DB bears higher equity value for barrier ratios below 0.5. Since the equity-maximizing DB is below 0.5 (as we have seen in Sects. 4.6.2 and 4.7), the equity holders will prefer triggering default according to a stochastic barrier. In fact this is a logical conclusion: the situation of the overall economy, as well as the size of the interest payments is taken into account.

5 Conclusion

This Chapter first provides a brief overview of the relevant structural models of asset pricing. This is followed by a discussion of the design of incentive compatible credit contracts in connection with game theory approach to the pricing of corporate assets. Finally these theoretical approaches were applied to the construction of a new Earnings Before Interest and Taxes (EBIT) based model of asset pricing.

[38] As it was discussed in Sect. 4.7, observable primitive variable implies low default triggering level. Consequently there is insignificant difference in the values produced by the two DB types.

The proposed structural model extends the available literature of asset pricing by an EBIT based model with stochastic interest rate. This framework is able to price equity and debt in a way consistent with the cash flow of the firm, and therefore to address some defects of the current frameworks. It solves the “delicate” issue of Leland [34], that the unlevered firm value might not be a traded asset, and deals with the problem of partial tax deductibility. The stochastic interest rate assumption contributes the possibility of analysing the effects of changes in the central bank’s monetary policy, and it is able to answer the question how the macroeconomical situation affects the optimal capital structure. The default is triggered using a stochastic default barrier, that is shown to be more accurate then its deterministic equivalent.

A weak point in our design is the assumption that the EBIT process is driven by a GBM, and therefore it cannot handle negative earnings. It might be argued that employing arithmetic Brownian motion would be a better choice for this reason, however it should be noted that our model has an infinite time horizon. As the prices of commodities grow exponentially, it is hard to accept a linear model for the EBIT evolution. Finding better alternatives for the EBIT process will be the subject of further research. A promising idea is to model the earnings as a difference of two correlated GBMs (representing revenues and expenses): it has a clear economic intuition, it is able to produce negative values, has an exponential expected evolution, and works with observable figures.

Acknowledgements The research leading to these results was supported by the European Union’s Horizon 2020 Research and Innovation Staff Exchange programme under the Marie Sklodowska-Curie grant agreement No 681228. The authors further acknowledge financial support from the Czech Science Foundation (grants number 16–00027S and 15–00036S). Karel Janda acknowledges research support provided during his long-term visits at Australian National University, Toulouse School of Economics, New Economic School, McGill University and University of California, Berkeley. The views expressed here are those of the authors and not necessarily those of our institutions. All remaining errors are solely our responsibility.

References

1. Altman, E.I., Brady, B., Resti, A., Sironi, A.: The link between default and recovery rates: theory, empirical evidence, and implications. J. Bus. **78**(6), 2203–2228 (2005)
2. Ang, J.S., Cole, R.A., Lin, J.W.: Agency costs and ownership structure. J. Financ. **55**(1), 81–106 (2000)
3. Barro, R.J.: The loan market, collateral, and rates of interest. J. Money Credit. Bank. **8**(4), 439–456 (1976)
4. Bebchuk, L.A.: Ex ante costs of violating absolute priority in bankruptcy. J. Financ. **57**(1), 445–460 (2002)
5. Bebchuk, L.A., Picker, R.C.: Bankruptcy rules, managerial entrenchment, and firm-specific human capital. In: Chicago Law and Economics Working Paper, vol.16 (1993)
6. Bernanke, B., Gertler, M.: Agency costs, net worth, and business fluctuations. Am. Econ. Rev. **79**(1), 14–31 (1989)
7. Bielecki, T.R., Rutkowski, M.: Credit Risk: Modeling, Valuation and Hedging. Springer, Berlin (2002)

8. Black, F., Cox, J.C.: Valuing corporate securities: some effects of bond indenture provisions. J. Financ. **31**(2), 351–367 (1976)
9. Black, F., Scholes, M.S.: The pricing of options and corporate liabilities. J. Politi. Econ. **81**(3), 637–654 (1973)
10. Bris, A., Welch, I., Zhu, N.: The costs of bankruptcy: chapter 7 liquidation versus chapter 11 reorganization. J. Financ. **61**(3), 1253–1303 (2006)
11. Briys, E., de Varenne, F.: Valuing fixed rate debt: an extention. J. Financ. Quant. Anal. **32**, 239–248 (1997)
12. Broadie, M., Chernov, M., Sundaresan, S.: Optimal debt and equity values in the presence of chapter 7 and chapter 11. J. Financ. **62**(3), 1341–1377 (2007)
13. Chan, Y.S., Kanatas, G.: Asymmetric valuations and the role of collateral in loan agreements. J. Money Credit. Bank. **17**(1), 84–95 (1985)
14. Collin-Dufresne, P., Goldstein, R.S.: Do credit spreads reflect stationary leverage ratios? J. Financ. **56**(5), 1929–1957 (2001)
15. Dozsa, M., Seidler, J.: Debt Contracts and Stochastic Default Barrier. In: Working Papers IES 2012/17, Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies (2012)
16. Duffie, D., Singleton, K.J.: Modeling term structures of defaultable bonds. Rev. Financ. Stud. **12**(4), 687–720 (1999)
17. Fischer, E.O., Heinkel, R., Zechner, J.: Dynamic capital structure choice: theory and tests. J. Financ. **44**(1), 19–40 (1989). Full publication date: Mar., 1989/Copyright 1989 American Finance Association
18. Franks, J.R., Torous, W.N.: An empirical investigation of US firms in reorganization. J. Financ. **44**(3), 747–769 (1989)
19. Gertner, R., Scharfstein, D.: A theory of workouts and the effects of reorganization law. J. Financ. **46**, 1189–1222 (1991)
20. Geske, R.: The valuation of corporate liabilities as compound options. J. Financ. Quant. Anal. **12**(04), 541–552 (1977)
21. Goldstein, R., Ju, N., Leland, H.: An EBIT-based model of dynamic capital structure. J. Bus. **74**(4), 483–512 (2001)
22. Graham, J.R.: How big are the tax benefits of debt? J. Financ. **55**(5), 1901–1941 (2000)
23. Hull, J., White, A.: The impact of default risk on the prices of options and other derivative securities. J. Bank. Financ. **19**(2), 299–322 (1995)
24. Ingersoll, J.E.: Theory of Financial Decision Making. Rowman & Littlefield, Totowa (1987)
25. Jackson, T.H.: Of liquidation, continuation, and delay: an analysis of bankruptcy policy and nonbankruptcy rules. Am. Bankr. LJ **60**, 399–428 (1986)
26. Janda, K.: Optimal debt contracts in emerging markets with multiple investors. Prague Econ. Pap. **16**(2), 115–129 (2007)
27. Janda, K.: Bankruptcies with soft budget constraint. Manch. School **77**(4), 430–460 (2009)
28. Janda, K., Rojcek, J.: Bankruptcy triggering asset value– continuous time finance approach. In: Pinto, A., Zilberman, D., (eds.) Modeling, Dynamics, Optimization and Bioeconomics I, in Springer Proceedings in Mathematics and Statistics, vol. 73, pp. 357–382. Springer (2014)
29. Jarrow, R.A., Lando, D., Turnbull, S.M.: A markov model for the term structure of credit risk spreads. Rev. Financ. Stud **10**(2), 481–523 (1997)
30. Jarrow, R.A., Turnbull, S.M.: Pricing derivatives on financial securities subject to credit risk. J. Financ. **50**(1), 53–85 (1995)
31. Kane, A., Marcus, A.J., McDonald, R.L.: How big is the tax advantage to debt? J. Financ. **39**(3), 841–853 (1984). Issue Title: Papers and Proceedings, Forty-Second Annual Meeting, American Finance Association, San Francisco, CA, December 28–30, 1983/Full publication date: Jul., 1984/Copyright 1984 American Finance Association
32. Kane, A., Marcus, A.J., Robert, L., Donald, M.C.: Debt policy and the rate of return premium to leverage. J. Financ. Quant. Anal. **20**(4), 479–499 (1985). Full publication date: Dec., 1985/Copyright 1985 University of Washington School of Business Administration
33. Lehrbass, F.: Defaulters get intense. Risk Credit. Risk Suppl. **10**(7), 56–59 (1997)

34. Leland, H.E.: Corporate debt value, bond covenants, and optimal capital structure. J. Financ. **49**(4), 1213–1252 (1994)
35. Leland, H.E., Toft, K.B.: Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. J. Financ. **51**(3), 987–1019 (1996)
36. Francis, A.: Longstaff and Eduardo S Schwartz. A simple approach to valuing risky fixed and floating rate debt. J. Financ. **50**(3), 789–819 (1995)
37. Madan, D.B., Unal, H.: Pricing the risks of default. Rev. Deriv. Res. **2**(2), 121–160 (1998)
38. Harry, M.: Markowitz. Portfolio selection. J. Financ. **7**(1), 77–91 (1952)
39. Meckling, W.H.: Financial markets, default, and bankruptcy: the role of the state. Law Contemp. Probs. **41**, 124–177 (1977)
40. Mella-Barral, P., Perraudin, W.: Strategic debt service. J. Financ. **52**(2), 531–556 (1997)
41. Mello, A.S., Parsons, J.E.: Measuring the agency cost of debt. J. Financ. **47**(5), 1887–1904 (1992)
42. Robert, C.: Merton. on the pricing of corporate debt: the risk structure of interest rates. J. Financ. **29**(2), 449–470 (1974)
43. Miller, M.H.: The wealth transfers of bankruptcy: some illustrative examples. Law. Contemp. Probl. **41**(4), 39–46 (1977)
44. Modigliani, F., Miller, M.H.: The cost of capital, corporation finance and the theory of investment. Am. Econ. Rev. **48**(3), 261–297 (1958)
45. Myers, S.C.: Determinants of corporate borrowing. J. Financ. Econ. **5**(2), 147–175 (1977)
46. Nejadmalayeri, A., Singh, M.: Corporate taxes, strategic default, and the cost of debt. J. Bank. Financ. **36**(11), 2900–2916 (2012)
47. Opler, T.C., Titman, S.: Financial distress and corporate performance. J. Financ. **49**(3), 1015–1040 (1994)
48. Povel, P.: Optimal'soft'or'tough'bankruptcy procedures. J. Law Econ. Organ. **15**(3), 659 (1999)
49. Reisel, N.: On the value of restrictive covenants: empirical investigation of public bond issues. J. Corp. Financ. **27**, 251–268 (2014)
50. Stiglitz, J.E., Weiss, A.: Credit rationing in markets with imperfect information. Am. Econ. Rev. **71**(3), 393–410 (1981)
51. Townsend, R.: Optimal contracts and competitive markets with costly state verification. J. Econ. Theory **21**(2), 265–293 (1979)
52. Vasicek, O.: An equilibrium characterization of the term structure. J. Financ. Econ. **5**(2), 177–188 (1977)
53. Vasicek, O.A.: Credit Valuation. KMV Corporation (1984)
54. Warner, J.B.: Bankruptcy, absolute priority, and the pricing of risky debt claims. J. Financ. Econ. **4**(3), 239–276 (1977)
55. Weiss, L.A.: Bankruptcy resolution: direct costs and violation of priority of claims. J. Financ. Econ. **27**(2), 285–314 (1990)
56. White, M.J.: The corporate bankruptcy decision. J. Econ. Perspect. **3**(2), 129–151 (1989)
57. Ziegler, A.: A Game Theory Analysis of Options: Corporate Finance and Financial Intermediation in Continous Time. Springer, Berlin (2004)

Direct and Inverse Variational Problems on Time Scales: A Survey

Monika Dryl and Delfim F.M. Torres

Abstract We deal with direct and inverse problems of the calculus of variations on arbitrary time scales. Firstly, using the Euler–Lagrange equation and the strengthened Legendre condition, we give a general form for a variational functional to attain a local minimum at a given point of the vector space. Furthermore, we provide a necessary condition for a dynamic integro-differential equation to be an Euler–Lagrange equation (Helmholtz's problem of the calculus of variations on time scales). New and interesting results for the discrete and quantum settings are obtained as particular cases. Finally, we consider very general problems of the calculus of variations given by the composition of a certain scalar function with delta and nabla integrals of a vector valued field.

Keywords Calculus of variations · Dynamic equations on time scales · Helmholtz's problem · Inverse problems · Self-adjoint equations · Equation of variation

1 Introduction

The theory of time scales is a relatively new area, which was introduced in 1988 by Stefan Hilger in his Ph.D. thesis [36–38]. It bridges, generalizes and extends the traditional discrete theory of dynamical systems (difference equations) and the theory for continuous dynamical systems (differential equations) [13] and the various dialects of q-calculus [29, 48] into a single unified theory [13, 14, 43].

The calculus of variations on time scales was introduced in 2004 by Martin Bohner [11] (see also [1, 39]) and has been developing rapidly in the past ten years, mostly due to its great potential for applications, e.g., in biology [13], economics [3, 7,

M. Dryl · D.F.M. Torres (✉)
Center for Research and Development in Mathematics and Applications (CIDMA),
Department of Mathematics, University of Aveiro, 3810–193 Aveiro, Portugal
e-mail: delfim@ua.pt

M. Dryl
e-mail: monikadryl@ua.pt

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_12

8, 31, 50] and mathematics [16, 32, 34, 62]. In order to deal with nontraditional applications in economics, where the system dynamics are described on a time scale partly continuous and partly discrete, or to accommodate nonuniform sampled systems, one needs to work with variational problems defined on a time scale [6, 8, 23, 27].

This survey is organized as follows. In Sect. 2 we review the basic notions of the time-scale calculus: the concepts of delta derivative and delta integral (Sect. 2.1); the analogous backward concepts of nabla differentiation and nabla integration (Sect. 2.2); and the relation between delta/forward and nabla/backward approaches (Sect. 2.3). Then, in Sect. 3, we review the central results of the recent and powerful calculus of variations on time scales. Both delta and nabla approaches are considered (Sects. 3.1 and 3.2, respectively). Our results begin with Sect. 4, where we investigate inverse problems of the calculus of variations on time scales. To our best knowledge, and in contrast with the direct problem, which is already well studied in the framework of time scales [62], the inverse problem has not been studied before. Its investigation is part of the Ph.D. thesis of the first author [23]. Let here, for the moment and just for simplicity, the time scale $\mathbb{T}$ be the set $\mathbb{R}$ of real numbers. Given L, a Lagrangian function, in the ordinary/direct fundamental problem of the calculus of variations one wants to find extremal curves $y : [a, b] \to \mathbb{R}^n$ giving stationary values to some action integral (functional)

$$\mathscr{I}(y) = \int_a^b L(t, y(t), y'(t))dt$$

with respect to variations of y with fixed boundary conditions $y(a) = y_a$ and $y(b) = y_b$. Thus, if in the direct problem we start with a Lagrangian and we end up with extremal curves, then one might expect as inverse problem to start with extremal curves and search for a Lagrangian. Such inverse problem is considered, in the general context of time scales, in Sect. 4.1: we describe a general form of a variational functional having an extremum at a given function y_0 under Euler–Lagrange and strengthened Legendre conditions (Theorem 16). In Corollary 2 the form of the Lagrangian L on the particular case of an isolated time scale is presented and we end Sect. 4.1 with some concrete cases and examples. We proceed with a more common inverse problem of the calculus of variations in Sect. 4.2. Indeed, normally the starting point are not the extremal curves but, instead, the Euler–Lagrange equations that such curves must satisfy:

$$\frac{\partial L}{\partial y} - \frac{d}{dt}\frac{\partial L}{\partial y'} = 0 \Leftrightarrow \frac{\partial L}{\partial y} - \frac{\partial^2 L}{\partial t \partial y'} - \frac{\partial^2 L}{\partial y \partial y'}y' - \frac{\partial^2 L}{\partial y' \partial y'}y'' = 0 \qquad (1)$$

(we are still keeping, for illustrative purposes, $\mathbb{T} = \mathbb{R}$). This is what is usually known as the inverse problem of the calculus of variations: start with a second order ordinary differential equation and determine a Lagrangian L (if it exists) whose Euler–Lagrange equations *are the same as* the given equation. The problem of variational formulation of differential equations (or the inverse problem of the calculus of variations) dates back to the 19th century. The problem seems to have been posed by Helmholtz in [35], followed by several results from [20, 21, 40, 54, 64]. There are, however, two different types of inverse problems, depending on the meaning of the phrase "are the same as". Do we require the equations to be the same or do we allow multiplication by functions to obtain new but equivalent equations? The first case is often called *Helmholtz's inverse problem*: find conditions under which a given differential equation is an Euler–Lagrange equation. The latter case is often called the *multiplier problem*: given $f(t, y, y', y'') = 0$, does a function $r(t, y, y')$ exist such that the equation $r(t, y, y')f(t, y, y', y'') = 0$ is the Euler–Lagrange equation of a functional? In this work we are interested in Helmholtz's problem. The answer to this problem in $\mathbb{T} = \mathbb{R}$ is classical and well known: the complete solution to Helmholtz's problem is found in the celebrated 1941 paper of Douglas [22]. Let O be a second order differential operator. Then, the differential equation $O(y) = 0$ is a second order Euler–Lagrange equation if and only if the Fréchet derivatives of O are self-adjoint. A simple example illustrating the difference between both inverse problems is the following one. Consider the second order differential equation

$$my'' + hy' + ky = f. \tag{2}$$

This equation is not self-adjoint and, as a consequence, there is no variational problem with such Euler–Lagrange equation. However, if we multiply the equation by $p(t) = \exp(ht/m)$, then

$$m\frac{d}{dt}\left[\exp(ht/m)y'\right] + k\ \exp(ht/m)y = \exp(ht/m)f \tag{3}$$

and now a variational formulation is possible: the Euler–Lagrange equation (1) associated with

$$\mathscr{I}(y) = \int_{t_0}^{t_1} \exp(ht/m)\left[\frac{1}{2}my'^2 - \frac{1}{2}ky^2 - fy\right]dt$$

is precisely (3). A recent theory of the calculus of variations that allows to obtain (2) directly has been developed, but involves Lagrangians depending on fractional (noninteger) order derivatives [4, 47, 49]. For a survey on the fractional calculus of variations, which is not our subject here, we refer the reader to [56]. For the time scale $\mathbb{T} = \mathbb{Z}$, available results on the inverse problem of the calculus of variations are more recent and scarcer. In this case Helmholtz's inverse problem can be formulated as follows: find conditions under which a second order difference equation is a second order discrete Euler–Lagrange equation. Available results in the literature go back to the works of Crăciun and Opriş (1996) and Albu and Opriş (1999) and, more recently, to the works of Hydon and Mansfeld (2004) and Bourdin and Cresson (2013) [2, 17, 19, 41]. The main difficulty to obtain analogous results to those of the classical continuous calculus of variations in the discrete (or, more generally, in the time-scale) setting is due to the lack of chain rule. This lack of chain rule is easily seen with a simple example. Let $f, g : \mathbb{Z} \to \mathbb{Z}$ be defined by $f(t) = t^3$, $g(t) = 2t$. Then, $\Delta(f \circ g)(t) = \Delta\left(8t^3\right) = 8\left(3t^2 + 3t + 1\right) = 24t^2 + 24t + 8$ and $\Delta f(g(t)) \cdot \Delta g(t) = \left(12t^2 + 6t + 1\right) 2 = 24t^2 + 12t + 2$. Therefore, $\Delta(f \circ g)(t) \neq \Delta f(g(t))\, \Delta g(t)$. The difficulties caused by the lack of a chain rule in a general time scale $\mathbb{T}$, in the context of the inverse problem of the calculus of variations on time scales, are discussed in Sect. 4.3. To deal with the problem, our approach to the inverse problem of the calculus of variations uses an integral perspective instead of the classical differential point of view. As a result, we obtain a useful tool to identify integro-differential equations which are not Euler–Lagrange equations on an arbitrary time scale $\mathbb{T}$. More precisely, we define the notion of self-adjointness of a first order integro-differential equation (Definition 15) and its equation of variation (Definition 16). Using such property, we prove a necessary condition for an integro-differential equation on an arbitrary time scale $\mathbb{T}$ to be an Euler–Lagrange equation (Theorem 17). In order to illustrate our results we present Theorem 17 in the particular time scales $\mathbb{T} \in \left\{\mathbb{R}, h\mathbb{Z}, \overline{q^{\mathbb{Z}}}\right\}$, $h > 0$, $q > 1$ (Corollaries 3–5). Furthermore, we discuss equivalences between: (i) integro-differential equations (20) and second order differential equations (29) (Proposition 1), and (ii) equations of variations of them on an arbitrary time scale $\mathbb{T}$ ((21) and (30), respectively). As a result, we show that it is impossible to prove the latter equivalence due to lack of a general chain rule on an arbitrary time scale [12, 13]. In Sect. 5 we address the direct problem of the calculus of variations on time scales by considering a variational problem which may be found often in economics (see [45] and references therein). We extremize a functional of the calculus of variations that is the composition of a certain scalar function with the delta and nabla integrals of a vector valued field, possibly subject to boundary conditions and/or isoperimetric constraints. In Sect. 5.1 we provide general Euler–Lagrange equations in integral form (Theorem 18), transversality conditions are given in Sect. 5.2, while Sect. 5.3 considers necessary optimality conditions for isoperimetric problems on an arbitrary time scale. Interesting corollaries and examples are presented in Sect. 5.4. We end with Sect. 6 of conclusions and open problems.

2 Preliminaries

A time scale $\mathbb{T}$ is an arbitrary nonempty closed subset of $\mathbb{R}$. The set of real numbers $\mathbb{R}$, the integers $\mathbb{Z}$, the natural numbers $\mathbb{N}$, the nonnegative integers $\mathbb{N}_0$, an union of closed intervals $[0, 1] \cup [2, 7]$ or the Cantor set are examples of time scales, while the set of rational numbers $\mathbb{Q}$, the irrational numbers $\mathbb{R} \setminus \mathbb{Q}$, the complex numbers $\mathbb{C}$ or an open interval like $(0, 1)$ are not time scales. Throughout this survey we assume that for $a, b \in \mathbb{T}$, $a < b$, all intervals are time scales intervals, i.e., $[a, b] = [a, b]_{\mathbb{T}} := [a, b] \cap \mathbb{T} = \{t \in \mathbb{T} : a \leq t \leq b\}$.

Table 1 Examples of jump operators and graininess functions on different time scales

$\mathbb{T}$	$\mathbb{R}$	$h\mathbb{Z}$	$\overline{q^{\mathbb{Z}}}$
$\sigma(t)$	t	$t+h$	qt
$\rho(t)$	t	$t-h$	$\frac{t}{q}$
$\mu(t)$	0	h	$t(q-1)$
$\nu(t)$	0	h	$\frac{t(q-1)}{q}$

Definition 1 (*See Sect. 1.1 of* [13]) Let $\mathbb{T}$ be a time scale and $t \in \mathbb{T}$. The forward jump operator $\sigma : \mathbb{T} \to \mathbb{T}$ is defined by $\sigma(t) := \inf\{s \in \mathbb{T} : s > t\}$ for $t \neq \sup \mathbb{T}$ and $\sigma(\sup \mathbb{T}) := \sup \mathbb{T}$ if $\sup \mathbb{T} < +\infty$. Accordingly, we define the backward jump operator $\rho : \mathbb{T} \to \mathbb{T}$ by $\rho(t) := \sup\{s \in \mathbb{T} : s < t\}$ for $t \neq \inf \mathbb{T}$ and $\rho(\inf \mathbb{T}) := \inf \mathbb{T}$ if $\inf \mathbb{T} > -\infty$. The forward graininess function $\mu : \mathbb{T} \to [0, \infty)$ is defined by $\mu(t) := \sigma(t) - t$ and the backward graininess function $\nu : \mathbb{T} \to [0, \infty)$ by $\nu(t) := t - \rho(t)$.

Example 1 The two classical time scales are $\mathbb{R}$ and $\mathbb{Z}$, representing the continuous and the purely discrete time, respectively. Other standard examples are the periodic numbers, $h\mathbb{Z} = \{hk : h > 0, k \in \mathbb{Z}\}$, and the q-scale

$$\overline{q^{\mathbb{Z}}} := q^{\mathbb{Z}} \cup \{0\} = \left\{q^k : q > 1, k \in \mathbb{Z}\right\} \cup \{0\}.$$

Sometimes one considers also the time scale $q^{\mathbb{N}_0} = \left\{q^k : q > 1, k \in \mathbb{N}_0\right\}$. The following time scale is common: $\mathbb{P}_{a,b} = \bigcup_{k=0}^{\infty} [k(a+b), k(a+b)+a]$, $a, b > 0$.

Table 1 and Example 2 present different forms of jump operators σ and ρ, and graininess functions μ and ν, in specified time scales.

Example 2 (See Example 1.2 of [65]) Let $a, b > 0$ and consider the time scale

$$\mathbb{P}_{a,b} = \bigcup_{k=0}^{\infty} [k(a+b), k(a+b)+a].$$

Then,

$$\sigma(t) = \begin{cases} t & \text{if } t \in A_1, \\ t+b & \text{if } t \in A_2, \end{cases} \qquad \rho(t) = \begin{cases} t-b & \text{if } t \in B_1, \\ t & \text{if } t \in B_2 \end{cases}$$

(see Fig. 1) and

$$\mu(t) = \begin{cases} 0 & \text{if } t \in A_1, \\ b & \text{if } t \in A_2, \end{cases} \qquad \nu(t) = \begin{cases} b & \text{if } t \in B_1, \\ 0 & \text{if } t \in B_2, \end{cases}$$

where

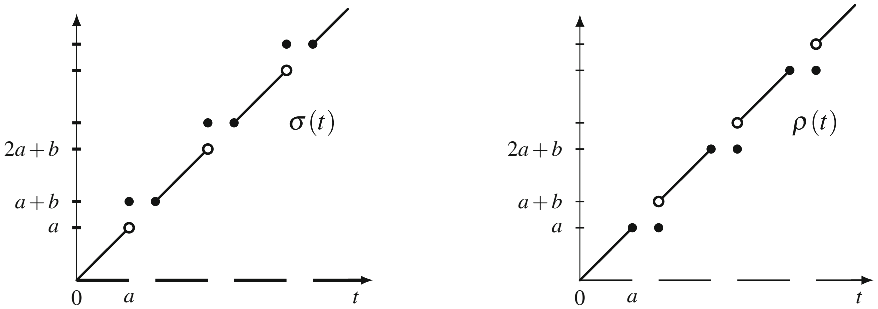


Fig. 1 Jump operators of the time scale $\mathbb{T} = \mathbb{P}_{a,b}$

$$\bigcup_{k=0}^{\infty}[k(a+b), k(a+b)+a] = A_1 \cup A_2 = B_1 \cup B_2$$

with

$$A_1 = \bigcup_{k=0}^{\infty}[k(a+b), k(a+b)+a), \quad B_1 = \bigcup_{k=0}^{\infty}\{k(a+b)\},$$

$$A_2 = \bigcup_{k=0}^{\infty}\{k(a+b)+a\}, \quad B_2 = \bigcup_{k=0}^{\infty}(k(a+b), k(a+b)+a].$$

In the time-scale theory the following classification of points is used:

- A point $t \in \mathbb{T}$ is called *right-scattered* or *left-scattered* if $\sigma(t) > t$ or $\rho(t) < t$, respectively.
- A point t is *isolated* if $\rho(t) < t < \sigma(t)$.
- If $t < \sup \mathbb{T}$ and $\sigma(t) = t$, then t is called *right-dense*; if $t > \inf \mathbb{T}$ and $\rho(t) = t$, then t is called *left-dense*.
- We say that t is *dense* if $\rho(t) = t = \sigma(t)$.

Definition 2 (*See Sect. 1 of* [55]) A time scale $\mathbb{T}$ is said to be an isolated time scale provided given any $t \in \mathbb{T}$, there is a $\delta > 0$ such that $(t-\delta, t+\delta) \cap \mathbb{T} = \{t\}$.

Definition 3 (*See* [9]) A time scale $\mathbb{T}$ is said to be regular if the following two conditions are satisfied simultaneously for all $t \in \mathbb{T}$: $\sigma(\rho(t)) = t$ and $\rho(\sigma(t)) = t$.

2.1 The Delta Derivative and the Delta Integral

If $f : \mathbb{T} \to \mathbb{R}$, then we define $f^{\sigma} : \mathbb{T} \to \mathbb{R}$ by $f^{\sigma}(t) := f(\sigma(t))$ for all $t \in \mathbb{T}$. The delta derivative (or *Hilger derivative*) of function $f : \mathbb{T} \to \mathbb{R}$ is defined for points in the set $\mathbb{T}^{\kappa}$, where

$$\mathbb{T}^\kappa := \begin{cases} \mathbb{T} \setminus \{\sup \mathbb{T}\} & \text{if } \rho(\sup \mathbb{T}) < \sup \mathbb{T} < \infty, \\ \mathbb{T} & \text{otherwise.} \end{cases}$$

Let us define the sets $\mathbb{T}^{\kappa^n}$, $n \geq 2$, inductively: $\mathbb{T}^{\kappa^1} := \mathbb{T}^\kappa$ and $\mathbb{T}^{\kappa^n} := \left(\mathbb{T}^{\kappa^{n-1}}\right)^\kappa$, $n \geq 2$. We define delta differentiability in the following way.

Definition 4 (*Sect. 1.1 of* [13]) Let $f : \mathbb{T} \to \mathbb{R}$ and $t \in \mathbb{T}^\kappa$. We define $f^\Delta(t)$ to be the number (provided it exists) with the property that given any $\varepsilon > 0$, there is a neighborhood U ($U = (t - \delta, t + \delta) \cap \mathbb{T}$ for some $\delta > 0$) of t such that

$$\left| f^\sigma(t) - f(s) - f^\Delta(t)\,(\sigma(t) - s) \right| \leq \varepsilon \left|\sigma(t) - s\right| \text{ for all } s \in U.$$

A function f is delta differentiable on $\mathbb{T}^\kappa$ provided $f^\Delta(t)$ exists for all $t \in \mathbb{T}^\kappa$. Then, $f^\Delta : \mathbb{T}^\kappa \to \mathbb{R}$ is called the delta derivative of f on $\mathbb{T}^\kappa$.

Theorem 1 (Theorem 1.16 of [13]) *Let $f : \mathbb{T} \to \mathbb{R}$ and $t \in \mathbb{T}^\kappa$. The following hold:*

1. *If f is delta differentiable at t, then f is continuous at t.*
2. *If f is continuous at t and t is right-scattered, then f is delta differentiable at t with*

$$f^\Delta(t) = \frac{f^\sigma(t) - f(t)}{\mu(t)}.$$

3. *If t is right-dense, then f is delta differentiable at t if and only if the limit*

$$\lim_{s \to t} \frac{f(t) - f(s)}{t - s}$$

exists as a finite number. In this case,

$$f^\Delta(t) = \lim_{s \to t} \frac{f(t) - f(s)}{t - s}.$$

4. *If f is delta differentiable at t, then*

$$f^\sigma(t) = f(t) + \mu(t) f^\Delta(t).$$

The next example is a consequence of Theorem 1 and presents different forms of the delta derivative on specific time scales.

Example 3 Let $\mathbb{T}$ be a time scale.

1. If $\mathbb{T} = \mathbb{R}$, then $f : \mathbb{R} \to \mathbb{R}$ is delta differentiable at $t \in \mathbb{R}$ if and only if

$$f^\Delta(t) = \lim_{s \to t} \frac{f(t) - f(s)}{t - s}$$

exists, i.e., if and only if f is differentiable (in the ordinary sense) at t and in this case we have $f^{\Delta}(t) = f'(t)$.

2. If $\mathbb{T} = h\mathbb{Z}$, $h > 0$, then $f : h\mathbb{Z} \to \mathbb{R}$ is delta differentiable at $t \in h\mathbb{Z}$ with

$$f^{\Delta}(t) = \frac{f(\sigma(t)) - f(t)}{\mu(t)} = \frac{f(t+h) - f(t)}{h} =: \Delta_h f(t).$$

In the particular case $h = 1$ we have $f^{\Delta}(t) = \Delta f(t)$, where Δ is the usual forward difference operator.

3. If $\mathbb{T} = \overline{q^{\mathbb{Z}}}$, $q > 1$, then for a delta differentiable function $f : \overline{q^{\mathbb{Z}}} \to \mathbb{R}$ we have

$$f^{\Delta}(t) = \frac{f(\sigma(t)) - f(t)}{\mu(t)} = \frac{f(qt) - f(t)}{(q-1)t} =: \Delta_q f(t)$$

for all $t \in \overline{q^{\mathbb{Z}}} \setminus \{0\}$, i.e., we get the usual Jackson derivative of quantum calculus [42, 48].

Now we formulate some basic properties of the delta derivative on time scales.

Theorem 2 (Theorem 1.20 of [13]) *Let $f, g : \mathbb{T} \to \mathbb{R}$ be delta differentiable at $t \in \mathbb{T}^{\kappa}$. Then,*

1. *the sum $f + g : \mathbb{T} \to \mathbb{R}$ is delta differentiable at t with*

$$(f+g)^{\Delta}(t) = f^{\Delta}(t) + g^{\Delta}(t);$$

2. *for any constant α, $\alpha f : \mathbb{T} \to \mathbb{R}$ is delta differentiable at t with*

$$(\alpha f)^{\Delta}(t) = \alpha f^{\Delta}(t);$$

3. *the product $fg : \mathbb{T} \to \mathbb{R}$ is delta differentiable at t with*

$$(fg)^{\Delta}(t) = f^{\Delta}(t)g(t) + f^{\sigma}(t)g^{\Delta}(t) = f(t)g^{\Delta}(t) + f^{\Delta}(t)g^{\sigma}(t);$$

4. *if $g(t)g^{\sigma}(t) \neq 0$, then f/g is delta differentiable at t with*

$$\left(\frac{f}{g}\right)^{\Delta}(t) = \frac{f^{\Delta}(t)g(t) - f(t)g^{\Delta}(t)}{g(t)g^{\sigma}(t)}.$$

Now we introduce the theory of delta integration on time scales. We start by defining the associated class of functions.

Definition 5 (*Sect. 1.4 of* [14]) A function $f : \mathbb{T} \to \mathbb{R}$ is called rd-continuous provided it is continuous at right-dense points in $\mathbb{T}$ and its left-sided limits exist (finite) at all left-dense points in $\mathbb{T}$.

The set of all rd-continuous functions $f : \mathbb{T} \to \mathbb{R}$ is denoted by $C_{rd} = C_{rd}(\mathbb{T}) = C_{rd}(\mathbb{T}, \mathbb{R})$. The set of functions $f : \mathbb{T} \to \mathbb{R}$ that are delta differentiable and whose derivative is rd-continuous is denoted by $C^1_{rd} = C^1_{rd}(\mathbb{T}) = C^1_{rd}(\mathbb{T}, \mathbb{R})$.

Definition 6 (*Definition 1.71 of* [13]) A function $F : \mathbb{T} \to \mathbb{R}$ is called a delta antiderivative of $f : \mathbb{T} \to \mathbb{R}$ provided $F^{\Delta}(t) = f(t)$ for all $t \in \mathbb{T}^{\kappa}$.

Definition 7 Let $\mathbb{T}$ be a time scale and $a, b \in \mathbb{T}$. If $f : \mathbb{T}^{\kappa} \to \mathbb{R}$ is a rd-continuous function and $F : \mathbb{T} \to \mathbb{R}$ is an antiderivative of f, then the Cauchy delta integral is defined by

$$\int_a^b f(t)\Delta t := F(b) - F(a).$$

Theorem 3 (Theorem 1.74 of [13]) *Every rd-continuous function f has an antiderivative F. In particular, if $t_0 \in \mathbb{T}$, then F defined by*

$$F(t) := \int_{t_0}^{t} f(\tau)\Delta\tau, \quad t \in \mathbb{T},$$

is an antiderivative of f.

Theorem 4 (Theorem 1.75 of [13]) *If $f \in C_{rd}$, then* $\displaystyle\int_t^{\sigma(t)} f(\tau)\Delta\tau = \mu(t)f(t)$, $t \in \mathbb{T}^{\kappa}$.

Let us see two special cases of the delta integral.

Example 4 Let $a, b \in \mathbb{T}$ and $f : \mathbb{T} \to \mathbb{R}$ be rd-continuous.

1. If $\mathbb{T} = \mathbb{R}$, then
$$\int_a^b f(t)\Delta t = \int_a^b f(t)dt,$$
where the integral on the right hand side is the usual Riemann integral.
2. If $[a, b]$ consists of only isolated points, then
$$\int_a^b f(t)\Delta t = \begin{cases} \sum\limits_{t\in[a,b)} \mu(t)f(t), & \text{if } a < b, \\ 0, & \text{if } a = b, \\ -\sum\limits_{t\in[b,a)} \mu(t)f(t), & \text{if } a > b. \end{cases}$$

Now we present some useful properties of the delta integral.

Theorem 5 (Theorem 1.77 of [13]) *If $a, b, c \in \mathbb{T}$, $a < c < b$, $\alpha \in \mathbb{R}$, and $f, g \in C_{rd}(\mathbb{T}, \mathbb{R})$, then:*

1. $\int_a^b (f(t) + g(t))\Delta t = \int_a^b f(t)\Delta t + \int_a^b g(t)\Delta t,$
2. $\int_a^b \alpha f(t)\Delta t = \alpha \int_a^b f(t)\Delta t,$
3. $\int_a^b f(t)\Delta t = -\int_b^a f(t)\Delta t,$
4. $\int_a^b f(t)\Delta t = \int_a^c f(t)\Delta t + \int_c^b f(t)\Delta t,$
5. $\int_a^a f(t)\Delta t = 0,$
6. $\int_a^b f(t)g^{\Delta}(t)\Delta t = f(t)g(t)|_{t=a}^{t=b} - \int_a^b f^{\Delta}(t)g^{\sigma}(t)\Delta t,$
7. $\int_a^b f^{\sigma}(t)g^{\Delta}(t)\Delta t = f(t)g(t)|_{t=a}^{t=b} - \int_a^b f^{\Delta}(t)g(t)\Delta t.$

2.2 The Nabla Derivative and the Nabla Integral

The nabla calculus is similar to the delta one of Sect. 2.1. The difference is that the backward jump operator ρ takes the role of the forward jump operator σ. For a function $f : \mathbb{T} \to \mathbb{R}$ we define $f^{\rho} : \mathbb{T} \to \mathbb{R}$ by $f^{\rho}(t) := f(\rho(t))$. If $\mathbb{T}$ has a right-scattered minimum m, then we define $\mathbb{T}_{\kappa} := \mathbb{T} - \{m\}$; otherwise, we set $\mathbb{T}_{\kappa} := \mathbb{T}$:

$$\mathbb{T}_{\kappa} := \begin{cases} \mathbb{T} \setminus \{\inf \mathbb{T}\} & \text{if } -\infty < \inf \mathbb{T} < \sigma(\inf \mathbb{T}), \\ \mathbb{T} & \text{otherwise.} \end{cases}$$

Let us define the sets $\mathbb{T}_{\kappa}$, $n \geq 2$, inductively: $\mathbb{T}_{\kappa^1} := \mathbb{T}_{\kappa}$ and $\mathbb{T}_{\kappa^n} := (\mathbb{T}_{\kappa^{n-1}})_{\kappa}$, $n \geq 2$. Finally, we define $\mathbb{T}_{\kappa}^{\kappa} := \mathbb{T}_{\kappa} \cap \mathbb{T}^{\kappa}$. The definition of nabla derivative of a function $f : \mathbb{T} \to \mathbb{R}$ at point $t \in \mathbb{T}_{\kappa}$ is similar to the delta case (cf. Definition 4).

Definition 8 (*Sect. 3.1 of* [14]) We say that a function $f : \mathbb{T} \to \mathbb{R}$ is nabla differentiable at $t \in \mathbb{T}_{\kappa}$ if there is a number $f^{\nabla}(t)$ such that for all $\varepsilon > 0$ there exists a neighborhood U of t (i.e., $U = (t - \delta, t + \delta) \cap \mathbb{T}$ for some $\delta > 0$) such that

$$|f^{\rho}(t) - f(s) - f^{\nabla}(t)(\rho(t) - s)| \leq \varepsilon|\rho(t) - s| \text{ for all } s \in U.$$

We say that $f^{\nabla}(t)$ is the nabla derivative of f at t. Moreover, f is said to be nabla differentiable on $\mathbb{T}$ provided $f^{\nabla}(t)$ exists for all $t \in \mathbb{T}_{\kappa}$.

The main properties of the nabla derivative are similar to those given in Theorems 1 and 2, and can be found, respectively, in Theorems 8.39 and 8.41 of [13].

Example 5 If $\mathbb{T} = \mathbb{R}$, then $f^{\nabla}(t) = f'(t)$. If $\mathbb{T} = h\mathbb{Z}$, $h > 0$, then

$$f^{\nabla}(t) = \frac{f(t) - f(t-h)}{h} =: \nabla_h f(t).$$

For $h = 1$ the operator ∇_h reduces to the standard backward difference operator $\nabla f(t) = f(t) - f(t-1)$.

We now briefly recall the theory of nabla integration on time scales. Similarly as in the delta case, first we define a suitable class of functions.

Definition 9 (*Sect. 3.1 of* [14]) Let $\mathbb{T}$ be a time scale and $f : \mathbb{T} \to \mathbb{R}$. We say that f is ld-continuous if it is continuous at left-dense points $t \in \mathbb{T}$ and its right-sided limits exist (finite) at all right-dense points.

Remark 1 If $\mathbb{T} = \mathbb{R}$, then f is ld-continuous if and only if f is continuous. If $\mathbb{T} = \mathbb{Z}$, then any function is ld-continuous.

The set of all ld-continuous functions $f : \mathbb{T} \to \mathbb{R}$ is denoted by $C_{ld} = C_{ld}(\mathbb{T}) = C_{ld}(\mathbb{T}, \mathbb{R})$; the set of all nabla differentiable functions with ld-continuous derivative by $C^1_{ld} = C^1_{ld}(\mathbb{T}) = C^1_{ld}(\mathbb{T}, \mathbb{R})$. Follows the definition of nabla integral on time scales.

Definition 10 (*Definition 8.42 of* [13]) A function $F : \mathbb{T} \to \mathbb{R}$ is called a nabla antiderivative of $f : \mathbb{T} \to \mathbb{R}$ provided $F^{\nabla}(t) = f(t)$ for all $t \in \mathbb{T}_{\kappa}$. In this case we define the nabla integral of f from a to b ($a, b \in \mathbb{T}$) by

$$\int_a^b f(t)\nabla t := F(b) - F(a).$$

Theorem 6 (Theorem 8.45 of [13] or Theorem 11 of [44]) *Every ld-continuous function f has a nabla antiderivative F. In particular, if $a \in \mathbb{T}$, then F defined by*

$$F(t) = \int_a^t f(\tau)\nabla\tau, \quad t \in \mathbb{T},$$

is a nabla antiderivative of f.

Theorem 7 (Theorem 8.46 of [13]) *If $f : \mathbb{T} \to \mathbb{R}$ is ld-continuous and $t \in \mathbb{T}_{\kappa}$, then*

$$\int_{\rho(t)}^t f(\tau)\nabla\tau = \nu(t) f(t).$$

Properties of the nabla integral, analogous to the ones of the delta integral given in Theorem 5, can be found in Theorem 8.47 of [13]. Here we give two special cases of the nabla integral.

Theorem 8 (See Theorem 8.48 of [13]) *Assume $a, b \in \mathbb{T}$ and $f : \mathbb{T} \to \mathbb{R}$ is ld-continuous.*

1. *If* $\mathbb{T} = \mathbb{R}$, *then*
$$\int_a^b f(t)\nabla t = \int_a^b f(t)dt,$$
where the integral on the right hand side is the Riemann integral.
2. *If* $\mathbb{T}$ *consists of only isolated points, then*
$$\int_a^b f(t)\nabla t = \begin{cases} \sum_{t\in(a,b]} \nu(t)f(t), & \textit{if } a < b, \\ 0, & \textit{if } a = b, \\ -\sum_{t\in(b,a]} \nu(t)f(t), & \textit{if } a > b. \end{cases}$$

2.3 Relation Between Delta and Nabla Operators

It is possible to relate the approach of Sect. 2.1 with that of Sect. 2.2.

Theorem 9 (See Theorems 2.5 and 2.6 of [5]) *If* $f : \mathbb{T} \to \mathbb{R}$ *is delta differentiable on* $\mathbb{T}^\kappa$ *and if* f^Δ *is continuous on* $\mathbb{T}^\kappa$*, then* f *is nabla differentiable on* $\mathbb{T}_\kappa$ *with*
$$f^\nabla(t) = \left(f^\Delta\right)^\rho(t) \textit{ for all } t \in \mathbb{T}_\kappa.$$
If $f : \mathbb{T} \to \mathbb{R}$ *is nabla differentiable on* $\mathbb{T}_\kappa$ *and if* f^∇ *is continuous on* $\mathbb{T}_\kappa$*, then* f *is delta differentiable on* $\mathbb{T}^\kappa$ *with*
$$f^\Delta(t) = \left(f^\nabla\right)^\sigma(t) \textit{ for all } t \in \mathbb{T}^\kappa. \tag{4}$$

Theorem 10 (Proposition 17 of [44]) *If function* $f : \mathbb{T} \to \mathbb{R}$ *is continuous, then for all* $a, b \in \mathbb{T}$ *with* $a < b$ *we have*
$$\int_a^b f(t)\Delta t = \int_a^b f^\rho(t)\nabla t,$$
$$\int_a^b f(t)\nabla t = \int_a^b f^\sigma(t)\Delta t.$$

For a more general theory relating delta and nabla approaches, we refer the reader to the duality theory of Caputo [18].

3 Direct Problems of the Calculus of Variations on Time Scales

There are two available approaches to the (direct) calculus of variations on time scales. The first one, the delta approach, is widely described in literature (see, e.g., [11, 13–15, 30, 31, 39, 44, 51, 62, 65]). The latter one, the nabla approach, was introduced mainly due to its applications in economics (see, e.g., [5–8]). It has been shown that these two types of calculus of variations are dual [18, 33, 46].

3.1 The Delta Approach to the Calculus of Variations

In this section we present the basic information about the delta calculus of variations on time scales. Let $\mathbb{T}$ be a given time scale with at least three points, and $a, b \in \mathbb{T}$, $a < b$, $a = \min \mathbb{T}$ and $b = \max \mathbb{T}$. Consider the following variational problem on the time scale $\mathbb{T}$:

$$\mathscr{L}[y] = \int\limits_a^b L\left(t, y^\sigma(t), y^\Delta(t)\right) \Delta t \longrightarrow \min \tag{5}$$

subject to the boundary conditions

$$y(a) = y_a, \quad y(b) = y_b, \quad y_a, y_b \in \mathbb{R}^n, \quad n \in \mathbb{N}. \tag{6}$$

Definition 11 A function $y \in C^1_{rd}(\mathbb{T}, \mathbb{R}^n)$ is said to be an admissible path (function) to problem (5)–(6) if it satisfies the given boundary conditions $y(a) = y_a$, $y(b) = y_b$.

In what follows the Lagrangian L is understood as a function $L : \mathbb{T} \times \mathbb{R}^{2n} \to \mathbb{R}$, $(t, y, v) \to L(t, y, v)$, and by L_y and L_v we denote the partial derivatives of L with respect to y and v, respectively. Similar notation is used for second order partial derivatives. We assume that $L(t, \cdot, \cdot)$ is differentiable in (y, v); $L(t, \cdot, \cdot)$, $L_y(t, \cdot, \cdot)$ and $L_v(t, \cdot, \cdot)$ are continuous at $\left(y^\sigma(t), y^\Delta(t)\right)$ uniformly at t and rd-continuous at t for any admissible path y. Let us consider the following norm in C^1_{rd}:

$$\|y\|_{C^1_{rd}} = \sup_{t \in [a,b]} \|y(t)\| + \sup_{t \in [a,b]^\kappa} \|y^\Delta(t)\|,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$.

Definition 12 We say that an admissible function $\hat{y} \in C^1_{rd}(\mathbb{T}; \mathbb{R}^n)$ is a local minimizer (respectively, a local maximizer) to problem (5)–(6) if there exists $\delta > 0$ such that $\mathscr{L}[\hat{y}] \leq \mathscr{L}[y]$ (respectively, $\mathscr{L}[\hat{y}] \geq \mathscr{L}[y]$) for all admissible functions $y \in C^1_{rd}(\mathbb{T}; \mathbb{R}^n)$ satisfying the inequality $||y - \hat{y}||_{C^1_{rd}} < \delta$.

Local minimizers (or maximizers) to problem (5)–(6) fulfill the delta differential Euler–Lagrange equation.

Theorem 11 (Delta differential Euler–Lagrange equation – see Theorem 4.2 of [11]) *If $\hat{y} \in C_{rd}^2(\mathbb{T}; \mathbb{R}^n)$ is a local minimizer to (5)–(6), then the Euler–Lagrange equation (in the delta differential form)*

$$L_v^{\Delta}\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right) = L_y\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right)$$

holds for $t \in [a, b]^{\kappa}$.

The next theorem provides the delta integral Euler–Lagrange equation.

Theorem 12 (Delta integral Euler–Lagrange equation – see [30, 39]) *If $\hat{y}(t) \in C_{rd}^1(\mathbb{T}; \mathbb{R}^n)$ is a local minimizer of the variational problem (5)–(6), then there exists a vector $c \in \mathbb{R}^n$ such that the Euler–Lagrange equation (in the delta integral form)*

$$L_v\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right) = \int\limits_a^t L_y(\tau, \hat{y}^{\sigma}(\tau), \hat{y}^{\Delta}(\tau))\Delta\tau + c^T \tag{7}$$

holds for $t \in [a, b]^{\kappa}$.

In the proof of Theorems 11 and 12 a time scale version of the Dubois–Reymond lemma is used.

Lemma 1 (See [11, 31]) *Let $f \in C_{rd}$, $f : [a, b] \to \mathbb{R}^n$. Then*

$$\int\limits_a^b f^T(t)\eta^{\Delta}(t)\Delta t = 0$$

holds for all $\eta \in C_{rd}^1([a, b], \mathbb{R}^n)$ with $\eta(a) = \eta(b) = 0$ if and only if $f(t) = c$ for all $t \in [a, b]^{\kappa}$, $c \in \mathbb{R}^n$.

The next theorem gives a second order necessary optimality condition for problem (5)–(6).

Theorem 13 (Legendre condition – see Result 1.3 of [11]) *If $\hat{y} \in C_{rd}^2(\mathbb{T}; \mathbb{R}^n)$ is a local minimizer of the variational problem (5)–(6), then*

$$A(t) + \mu(t)\left\{C(t) + C^T(t) + \mu(t)B(t) + (\mu(\sigma(t)))^{\dagger}A(\sigma(t))\right\} \geq 0, \tag{8}$$

$t \in [a, b]^{\kappa^2}$, where

$$\begin{aligned} A(t) &= L_{vv}\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right), \\ B(t) &= L_{yy}\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right), \\ C(t) &= L_{yv}\left(t, \hat{y}^{\sigma}(t), \hat{y}^{\Delta}(t)\right) \end{aligned}$$

and where $\alpha^{\dagger} = \frac{1}{\alpha}$ *if* $\alpha \in \mathbb{R} \setminus \{0\}$ *and* $0^{\dagger} = 0$.

Remark 2 If (8) holds with the strict inequality "$>$", then it is called *the strengthened Legendre condition*.

3.2 The Nabla Approach to the Calculus of Variations

In this section we consider a problem of the calculus of variations that involves a functional with a nabla derivative and a nabla integral. The motivation to study such variational problems is coming from applications, in particular from economics [7, 8]. Let $\mathbb{T}$ be a given time scale, which has sufficiently many points in order for all calculations to make sense, and let $a, b \in \mathbb{T}$, $a < b$. The problem consists of minimizing or maximizing

$$\mathscr{L}[y] = \int_{a}^{b} L\left(t, y^{\rho}(t), y^{\nabla}(t)\right) \nabla t \tag{9}$$

in the class of functions $y \in C_{ld}^{1}(\mathbb{T}; \mathbb{R}^n)$ subject to the boundary conditions

$$y(a) = y_a, \quad y(b) = y_b, \quad y_a, y_b \in \mathbb{R}^n, \quad n \in \mathbb{N}. \tag{10}$$

Definition 13 A function $y \in C_{ld}^{1}(\mathbb{T}, \mathbb{R}^n)$ is said to be an admissible path (function) to problem (9)–(10) if it satisfies the given boundary conditions $y(a)=y_a$, $y(b) = y_b$.

As before, the Lagrangian L is understood as a function $L : \mathbb{T} \times \mathbb{R}^{2n} \to \mathbb{R}$, $(t, y, v) \to L(t, y, v)$. We assume that $L(t, \cdot, \cdot)$ is differentiable in (y, v); $L(t, \cdot, \cdot)$, $L_y(t, \cdot, \cdot)$ and $L_v(t, \cdot, \cdot)$ are continuous at $\left(y^{\rho}(t), y^{\nabla}(t)\right)$ uniformly at t and ld-continuous at t for any admissible path y. Let us consider the following norm in C_{ld}^{1}:

$$\|y\|_{C_{ld}^{1}} = \sup_{t \in [a,b]} \|y(t)\| + \sup_{t \in [a,b]_{\kappa}} \|y^{\nabla}(t)\|$$

with $\| \cdot \|$ the Euclidean norm in $\mathbb{R}^n$.

Definition 14 (*See* [3]) We say that an admissible function $y \in C_{ld}^{1}(\mathbb{T}; \mathbb{R}^n)$ is a local minimizer (respectively, a local maximizer) for the variational problem (9)–(10) if there exists $\delta > 0$ such that $\mathscr{L}[\hat{y}] \leq \mathscr{L}[y]$ (respectively, $\mathscr{L}[\hat{y}] \geq \mathscr{L}[y]$) for all $y \in C_{ld}^{1}(\mathbb{T}; \mathbb{R}^n)$ satisfying the inequality $||y - \hat{y}||_{C_{ld}^{1}} < \delta$.

In case of the first order necessary optimality condition for nabla variational problem on time scales (9)–(10), the Euler–Lagrange equation takes the following form.

Theorem 14 (Nabla Euler–Lagrange equation – see [62]) *If a function* $\hat{y} \in C_{ld}^1(\mathbb{T}; \mathbb{R}^n)$ *provides a local extremum to the variational problem* (9)–(10)*, then* $\hat{y}$ *satisfies the Euler–Lagrange equation (in the nabla differential form)*

$$L_v^\nabla\left(t, y^\rho(t), y^\nabla(t)\right) = L_y\left(t, y^\rho(t), y^\nabla(t)\right)$$

for all $t \in [a, b]_\kappa$.

Now we present the fundamental lemma of the nabla calculus of variations on time scales.

Lemma 2 (See [50]) *Let* $f \in C_{ld}([a, b], \mathbb{R}^n)$*. If*

$$\int\limits_a^b f(t)\eta^\nabla(t)\nabla t = 0$$

for all $\eta \in C_{ld}^1([a, b], \mathbb{R}^n)$ *with* $\eta(a) = \eta(b) = 0$*, then* $f(t) = c$ *for all* $t \in [a, b]_\kappa$, $c \in \mathbb{R}^n$.

For a good survey on the direct calculus of variations on time scales, covering both delta and nabla approaches, we refer the reader to [62].

4 Inverse Problems of the Calculus of Variations on Time Scales

This section is devoted to inverse problems of the calculus of variations on an arbitrary time scale. To our best knowledge, the inverse problem has not been studied before 2014 [23, 26, 28] in the framework of time scales, in contrast with the direct problem, that establishes dynamic equations of Euler–Lagrange type to time-scale variational problems, that has now been investigated for ten years, since 2004 [11]. To begin (Sect. 4.1) we consider an inverse extremal problem associated with the following fundamental problem of the calculus of variations: to minimize

$$\mathscr{L}[y] = \int\limits_a^b L\left(t, y^\sigma(t), y^\Delta(t)\right)\Delta t \tag{11}$$

subject to boundary conditions $y(a) = y_0(a)$, $y(b) = y_0(b)$ on a given time scale $\mathbb{T}$. The Euler–Lagrange equation and the strengthened Legendre condition are used in order to describe a general form of a variational functional (11) that attains an extremum at a given function y_0. In the latter Sect. 4.2, we introduce a completely different approach to the inverse problem of the calculus of variations, using an integral perspective instead of the classical differential point of view [17, 21]. We present a sufficient condition of self-adjointness for an integro-differential equation (Lemma 3). Using this property, we prove a necessary condition for an integro-differential equation on an arbitrary time scale $\mathbb{T}$ to be an Euler–Lagrange equation (Theorem 17), related to a property of self-adjointness (Definition 15) of its equation of variation (Definition 16).

4.1 A General Form of the Lagrangian

The problem under our consideration is to find a general form of the variational functional

$$\mathscr{L}[y] = \int_a^b L\left(t, y^\sigma(t), y^\Delta(t)\right) \Delta t \tag{12}$$

subject to boundary conditions $y(a) = y(b) = 0$, possessing a local minimum at zero, under the Euler–Lagrange and the strengthened Legendre conditions. We assume that $L(t, \cdot, \cdot)$ is a C^2-function with respect to (y, v) uniformly in t, and L, L_y, L_v, $L_{vv} \in C_{rd}$ for any admissible path $y(\cdot)$. Observe that under our assumptions, by Taylor's theorem, we may write L, with the big O notation, in the form

$$L(t, y, v) = P(t, y) + Q(t, y)v + \frac{1}{2}R(t, y, 0)v^2 + O(v^3), \tag{13}$$

where

$$\begin{aligned} P(t, y) &= L(t, y, 0), \\ Q(t, y) &= L_v(t, y, 0), \\ R(t, y, 0) &= L_{vv}(t, y, 0). \end{aligned} \tag{14}$$

Let $R(t, y, v) = R(t, y, 0) + O(v)$. Then, one can write (13) as

$$L(t, y, v) = P(t, y) + Q(t, y)v + \frac{1}{2}R(t, y, v)v^2.$$

Now the idea is to find general forms of $P(t, y^\sigma(t))$, $Q(t, y^\sigma(t))$ and $R(t, y^\sigma(t), y^\Delta(t))$ using the Euler–Lagrange (7) and the strengthened Legendre (8) conditions with notation (14). Then we use the Euler–Lagrange equation (7) and choose an arbitrary function $P(t, y^\sigma(t))$ such that $P(t, \cdot) \in C^2$ with respect to the second variable, uniformly in t, P and P_y rd-continuous in t for all admissible y. We can write the general form of Q as

$$Q(t, y^\sigma(t)) = C + \int_a^t P_y(\tau, 0)\Delta\tau + q(t, y^\sigma(t)) - q(t, 0),$$

where $C \in \mathbb{R}$ and q is an arbitrarily function such that $q(t, \cdot) \in C^2$ with respect to the second variable, uniformly in t, q and q_y are rd-continuous in t for all admissible y. From the strengthened Legendre condition (8), with notation (14), we set

$$R(t, 0, 0) + \mu(t)\left\{2Q_y(t, 0) + \mu(t)P_{yy}(t, 0) + (\mu^\sigma(t))^\dagger R(\sigma(t), 0, 0)\right\} = p(t) \tag{15}$$

with $p \in C_{rd}([a, b])$, $p(t) > 0$ for all $t \in [a, b]^{\kappa^2}$, chosen arbitrary, where $\alpha^\dagger = \frac{1}{\alpha}$ if $\alpha \in \mathbb{R} \setminus \{0\}$ and $0^\dagger = 0$. We obtain the following theorem, which presents a general form of the integrand L for functional (12).

Theorem 15 *Let $\mathbb{T}$ be an arbitrary time scale. If functional (12) with boundary conditions $y(a) = y(b) = 0$ attains a local minimum at $\hat{y}(t) \equiv 0$ under the strengthened Legendre condition, then its Lagrangian L takes the form*

$$\begin{aligned} L\left(t, y^\sigma(t), y^\Delta(t)\right) &= P\left(t, y^\sigma(t)\right) \\ &+ \left(C + \int_a^t P_y(\tau, 0)\Delta\tau + q(t, y^\sigma(t)) - q(t, 0)\right) y^\Delta(t) \\ &+ \Bigg(p(t) - \mu(t)\left\{2Q_y(t, 0) + \mu(t)P_{yy}(t, 0) + (\mu^\sigma(t))^\dagger R(\sigma(t), 0, 0)\right\} \\ &\quad + w(t, y^\sigma(t), y^\Delta(t)) - w(t, 0, 0)\Bigg)\frac{(y^\Delta(t))^2}{2}, \end{aligned}$$

where $R(t, 0, 0)$ is a solution of equation (15), $C \in \mathbb{R}$, $\alpha^\dagger = \frac{1}{\alpha}$ if $\alpha \in \mathbb{R} \setminus \{0\}$ and $0^\dagger = 0$. Functions P, p, q and w are arbitrary functions satisfying:

(i) $P(t, \cdot), q(t, \cdot) \in C^2$ with respect to the second variable uniformly in t; P, P_y, q, q_y are rd-continuous in t for all admissible y; $P_{yy}(\cdot, 0)$ is rd-continuous in t; $p \in C^1_{rd}$ with $p(t) > 0$ for all $t \in [a, b]^{\kappa^2}$;

(ii) $w(t, \cdot, \cdot) \in C^2$ with respect to the second and the third variable, uniformly in t; w, w_y, w_v, w_{vv} are rd-continuous in t for all admissible y.

Proof See [28].

Now we consider the general situation when the variational problem consists in minimizing (12) subject to arbitrary boundary conditions $y(a) = y_0(a)$ and $y(b) = y_0(b)$, for a certain given function $y_0 \in C^2_{rd}([a, b])$.

Theorem 16 *Let $\mathbb{T}$ be an arbitrary time scale. If the variational functional (12) with boundary conditions $y(a) = y_0(a)$, $y(b) = y_0(b)$, attains a local minimum for a certain given function $y_0(\cdot) \in C^2_{rd}([a, b])$ under the strengthened Legendre condition, then its Lagrangian L has the form*

$$\begin{aligned}
&L\left(t, y^\sigma(t), y^\Delta(t)\right) = P\left(t, y^\sigma(t) - y_0^\sigma(t)\right) + \left(y^\Delta(t) - y_0^\Delta(t)\right)\\
&\times\left(C + \int_a^t P_y\left(\tau, -y_0^\sigma(\tau)\right)\Delta\tau + q\left(t, y^\sigma(t) - y_0^\sigma(t)\right) - q\left(t, -y_0^\sigma(t)\right)\right) + \frac{1}{2}\Bigg(p(t)\\
&- \mu(t)\left\{2Q_y(t, -y_0^\sigma(t)) + \mu(t)P_{yy}(t, -y_0^\sigma(t)) + \left(\mu^\sigma(t)\right)^\dagger R(\sigma(t), -y_0^\sigma(t), -y_0^\Delta(t))\right\}\\
&+ w(t, y^\sigma(t) - y_0^\sigma(t), y^\Delta(t) - y_0^\Delta(t)) - w\left(t, -y_0^\sigma(t), -y_0^\Delta(t)\right)\Bigg)\left(y^\Delta(t) - y_0^\Delta(t)\right)^2,
\end{aligned}$$

where $C \in \mathbb{R}$ and functions $P,\ p,\ q,\ w$ satisfy conditions (i) and (ii) of Theorem 15.

Proof See [28].

For the classical situation $\mathbb{T} = \mathbb{R}$, Theorem 16 gives a recent result of [57, 58].

Corollary 1 (Theorem 4 of [57]) *If the variational functional*

$$\mathscr{L}[y] = \int_a^b L(t, y(t), y'(t))dt$$

attains a local minimum at $y_0(\cdot) \in C^2[a, b]$ when subject to boundary conditions $y(a) = y_0(a)$ and $y(b) = y_0(b)$ and the classical strengthened Legendre condition

$$R(t, y_0(t), y_0'(t)) > 0, \quad t \in [a, b],$$

then its Lagrangian L has the form

$$\begin{aligned}
&L(t, y(t), y'(t)) = P(t, y(t) - y_0(t))\\
&+ (y'(t) - y_0'(t))\left(C + \int_a^t P_y(\tau, -y_0(\tau))d\tau + q(t, y(t) - y_0(t)) - q(t, -y_0(t))\right)\\
&+ \frac{1}{2}\left(p(t) + w(t, y(t) - y_0(t), y'(t) - y_0'(t)) - w(t, -y_0(t), -y_0'(t))\right)(y'(t) - y_0'(t))^2,
\end{aligned}$$

where $C \in \mathbb{R}$.

In the particular case of an isolated time scale, where $\mu(t) \neq 0$ for all $t \in \mathbb{T}$, we get the following corollary.

Corollary 2 *Let $\mathbb{T}$ be an isolated time scale. If functional (12) subject to the boundary conditions $y(a) = y(b) = 0$ attains a local minimum at $\hat{y}(t) \equiv 0$ under the strengthened Legendre condition, then the Lagrangian L has the form*

$$L\left(t, y^{\sigma}(t), y^{\Delta}(t)\right) = P\left(t, y^{\sigma}(t)\right)$$
$$+\left(C + \int_a^t P_y(\tau, 0)\Delta\tau + q(t, y^{\sigma}(t)) - q(t, 0)\right) y^{\Delta}(t)$$
$$+\left(e_r(t, a)R_0 + \int_a^t e_r(t, \sigma(\tau))s(\tau)\Delta\tau + w(t, y^{\sigma}(t), y^{\Delta}(t)) - w(t, 0, 0)\right) \frac{(y^{\Delta}(t))^2}{2},$$

where $C, R_0 \in \mathbb{R}$ and $r(t)$ and $s(t)$ are given by

$$r(t) := -\frac{1 + \mu(t)(\mu^{\sigma}(t))^{\dagger}}{\mu^2(t)(\mu^{\sigma}(t))^{\dagger}}, \quad s(t) := \frac{p(t) - \mu(t)[2Q_y(t, 0) + \mu(t)P_{yy}(t, 0)]}{\mu^2(t)(\mu^{\sigma}(t))^{\dagger}} \tag{16}$$

with $\alpha^{\dagger} = \frac{1}{\alpha}$ if $\alpha \in \mathbb{R} \setminus \{0\}$ and $0^{\dagger} = 0$, and functions P, p, q, w satisfy assumptions of Theorem 15.

Based on Corollary 2, we present the form of Lagrangian L in the periodic time scale $\mathbb{T} = h\mathbb{Z}$.

Example 6 Let $\mathbb{T} = h\mathbb{Z}$, $h > 0$, and $a, b \in h\mathbb{Z}$ with $a < b$. Then $\mu(t) \equiv h$. Consider the variational functional

$$\mathscr{L}[y] = h \sum_{k=\frac{a}{h}}^{\frac{b}{h}-1} L\left(kh, y(kh + h), \Delta_h y(kh)\right) \tag{17}$$

subject to the boundary conditions $y(a) = y(b) = 0$, which attains a local minimum at $\hat{y}(kh) \equiv 0$ under the strengthened Legendre condition

$$R(kh, 0, 0) + 2hQ_y(kh, 0) + h^2 P_{yy}(kh, 0) + R(kh + h, 0, 0) > 0,$$

$kh \in [a, b - 2h] \cap h\mathbb{Z}$. Functions $r(t)$ and $s(t)$ (see (16)) have the following form:

$$r(t) \equiv -\frac{2}{h}, \quad s(t) = \frac{p(t)}{h} - \left(2Q_y(t, 0) + hP_{yy}(t, 0)\right).$$

Hence,

$$\int_a^t P_y(\tau, 0)\Delta\tau = h \sum_{i=\frac{a}{h}}^{\frac{t}{h}-1} P_y(ih, 0),$$

$$\int_a^t e_r(t, \sigma(\tau))s(\tau)\Delta\tau = \sum_{i=\frac{a}{h}}^{\frac{t}{h}-1} (-1)^{\frac{t}{h}-i-1} \left(p(ih) - 2hQ_y(ih, 0) - h^2 P_{yy}(ih, 0)\right).$$

Thus, the Lagrangian L of the variational functional (17) on $\mathbb{T} = h\mathbb{Z}$ has the form

$$\begin{aligned} L\left(kh, y(kh+h), \Delta_h y(kh)\right) &= P\left(kh, y(kh+h)\right) \\ &+ \left(C + \sum_{i=\frac{a}{h}}^{k-1} hP_y(ih,0) + q(kh, y(kh+h)) - q(kh,0)\right)\Delta_h y(kh) \\ &+ \frac{1}{2}\Bigg((-1)^{k-\frac{a}{h}} R_0 + \sum_{i=\frac{a}{h}}^{k-1} (-1)^{k-i-1}\left(p(ih) - 2hQ_y(ih,0) - h^2 P_{yy}(ih,0)\right) \\ &\quad + w(kh, y(kh+h), \Delta_h y(kh)) - w(kh,0,0)\Bigg)\left(\Delta_h y(kh)\right)^2, \end{aligned}$$

where functions P, p, q, w are arbitrary but satisfy assumptions of Theorem 15.

4.2 Necessary Condition for an Euler–Lagrange Equation

This section provides a necessary condition for an integro-differential equation on an arbitrary time scale to be an Euler–Lagrange equation (Theorem 17). For that the notions of self-adjointness (Definition 15) and equation of variation (Definition 16) are essential.

Definition 15 (*First order self-adjoint integro-differential equation*) A first order integro-differential dynamic equation is said to be *self-adjoint* if it has the form

$$Lu(t) = const, \text{ where } Lu(t) = p(t)u^{\Delta}(t) + \int_{t_0}^{t} \left[r(s)u^{\sigma}(s)\right]\Delta s \tag{18}$$

with $p, r \in C_{rd}$, $p \neq 0$ for all $t \in \mathbb{T}$ and $t_0 \in \mathbb{T}$.

Let $\mathbb{D}$ be the set of all functions $y : \mathbb{T} \to \mathbb{R}$ such that $y^{\Delta} : \mathbb{T}^{\kappa} \to \mathbb{R}$ is continuous. A function $y \in \mathbb{D}$ is said to be a solution of (18) provided $Ly(t) = const$ holds for all $t \in \mathbb{T}^{\kappa}$. For simplicity, we use the operators $[\cdot]$ and $\langle\cdot\rangle$ defined as

$$[y](t) := (t, y^{\sigma}(t), y^{\Delta}(t)), \qquad \langle y\rangle(t) := (t, y^{\sigma}(t), y^{\Delta}(t), y^{\Delta\Delta}(t)), \tag{19}$$

and partial derivatives of function $(t, y, v, z) \to L(t, y, v, z)$ are denoted by $\partial_2 L = L_y$, $\partial_3 L = L_v$, $\partial_4 L = L_z$.

Definition 16 (*Equation of variation*) Let

$$H[y](t) + \int_{t_0}^{t} G[y](s)\Delta s = const \tag{20}$$

be an integro-differential equation on time scales with $H_v \neq 0$, $t \to F_y[y](t)$, $t \to F_v[y](t) \in C_{rd}(\mathbb{T}, \mathbb{R})$ along every curve y, where $F \in \{G, H\}$. The *equation of variation* associated with (20) is given by

$$H_y[u](t)u^\sigma(t) + H_v[u](t)u^\Delta(t) + \int_{t_0}^{t} G_y[u](s)u^\sigma(s) + G_v[u](s)u^\Delta(s)\Delta s = 0. \tag{21}$$

Lemma 3 (Sufficient condition of self-adjointness) *Let* (20) *be a given integro-differential equation. If*

$$H_y[y](t) + G_v[y](t) = 0,$$

then its equation of variation (21) *is self-adjoint.*

Proof See [26].

Now we provide an answer to the general inverse problem of the calculus of variations on time scales.

Theorem 17 (Necessary condition for an Euler–Lagrange equation in integral form) *Let* $\mathbb{T}$ *be an arbitrary time scale and*

$$H(t, y^\sigma(t), y^\Delta(t)) + \int_{t_0}^{t} G(s, y^\sigma(s), y^\Delta(s))\Delta s = const \tag{22}$$

be a given integro-differential equation. If (22) *is to be an Euler–Lagrange equation, then its equation of variation* (21) *is self-adjoint in the sense of Definition 15.*

Proof See [26].

Remark 3 In practical terms, Theorem 17 is useful to identify equations that are not Euler–Lagrange: if the equation of variation (21) of a given dynamic equation (20) is not self-adjoint, then we conclude that (20) is not an Euler–Lagrange equation.

Now we present an example of a second order differential equation on time scales which is not an Euler–Lagrange equation.

Example 7 Let us consider the following second-order linear oscillator dynamic equation on an arbitrary time scale $\mathbb{T}$:

$$y^{\Delta\Delta}(t) + y^\Delta(t) - t = 0. \tag{23}$$

We may write Eq. (23) in integro-differential form (20):

$$y^\Delta(t) + \int_{t_0}^{t} \left(y^\Delta(s) - s\right) \Delta s = const, \tag{24}$$

where $H[y](t) = y^{\Delta}(t)$ and $G[y](t) = y^{\Delta}(t) - t$. Because

$$H_y[y](t) = G_y[y](t) = 0, \quad H_v[y](t) = G_v[y](t) = 1,$$

the equation of variation associated with (24) is given by

$$u^{\Delta}(t) + \int_{t_0}^{t} u^{\Delta}(s)\Delta s = 0 \iff u^{\Delta}(t) + u(t) = u(t_0). \tag{25}$$

We may notice that Eq. (25) cannot be written in form (18), hence, it is not self-adjoint. Following Theorem 17 (see Remark 3) we conclude that Eq. (23) is not an Euler–Lagrange equation.

Now we consider the particular case of Theorem 17 when $\mathbb{T} = \mathbb{R}$ and $y \in C^2([t_0, t_1]; \mathbb{R})$. In this case operator $[\cdot]$ of (19) has the form

$$[y](t) = (t, y(t), y'(t)) =: [y]_{\mathbb{R}}(t),$$

while condition (18) can be written as

$$p(t)u'(t) + \int_{t_0}^{t} r(s)u(s)ds = const. \tag{26}$$

Corollary 3 *If a given integro-differential equation*

$$H(t, y(t), y'(t)) + \int_{t_0}^{t} G(s, y(s), y'(s))ds = const$$

is to be the Euler–Lagrange equation of the variational problem

$$\mathscr{I}[y] = \int_{t_0}^{t_1} L(t, y(t), y'(t))dt$$

(cf., e.g., [63]), then its equation of variation

$$H_y[u]_{\mathbb{R}}(t)u(t) + H_v[u]_{\mathbb{R}}(t)u'(t) + \int_{t_0}^{t} G_y[u]_{\mathbb{R}}(s)u(s) + G_v[u]_{\mathbb{R}}(s)u'(s)ds = 0$$

must be self-adjoint, in the sense of Definition 15 with (18) *given by* (26).

Proof Follows from Theorem 17 with $\mathbb{T} = \mathbb{R}$.

Now we consider the particular case of Theorem 17 when $\mathbb{T} = h\mathbb{Z}$, $h > 0$. In this case operator $[\cdot]$ of (19) has the form

$$[y](t) = (t, y(t+h), \Delta_h y(t)) =: [y]_h(t),$$

where

$$\Delta_h y(t) = \frac{y(t+h) - y(t)}{h}.$$

For $\mathbb{T} = h\mathbb{Z}$, $h > 0$, condition (18) can be written as

$$p(t)\Delta_h u(t) + \sum_{k=\frac{t_0}{h}}^{\frac{t}{h}-1} hr(kh)u(kh+h) = const. \tag{27}$$

Corollary 4 *If a given difference equation*

$$H(t, y(t+h), \Delta_h y(t)) + \sum_{k=\frac{t_0}{h}}^{\frac{t}{h}-1} hG(kh, y(kh+h), \Delta_h y(kh)) = const$$

is to be the Euler–Lagrange equation of the discrete variational problem

$$\mathscr{I}[y] = \sum_{k=\frac{t_0}{h}}^{\frac{t_1}{h}-1} hL(kh, y(kh+h), \Delta_h y(kh))$$

(cf., e.g., [10]), then its equation of variation

$$\begin{aligned} &H_y[u]_h(t)u(t+h) + H_v[u]_h(t)\Delta_h u(t) \\ &\qquad + h\sum_{k=\frac{t_0}{h}}^{\frac{t}{h}-1} \big(G_y[u]_h(kh)u(kh+h) + G_v[u]_h(kh)\Delta_h u(kh)\big) = 0 \end{aligned}$$

is self-adjoint, in the sense of Definition 15 with (18) *given by* (27).

Proof Follows from Theorem 17 with $\mathbb{T} = h\mathbb{Z}$.

Finally, let us consider the particular case of Theorem 17 when $\mathbb{T} = \overline{q^{\mathbb{Z}}} = q^{\mathbb{Z}} \cup \{0\}$, where $q^{\mathbb{Z}} = \left\{q^k : k \in \mathbb{Z}, q > 1\right\}$. In this case operator $[\cdot]$ of (19) has the form

$$[y]_{\overline{q^{\mathbb{Z}}}}(t) = (t, y(qt), \Delta_q y(t)) =: [y]_q(t),$$

where

$$\Delta_q y(t) = \frac{y(qt) - y(t)}{(q-1)t}.$$

For $\mathbb{T} = \overline{q^{\mathbb{Z}}}$, $q > 1$, condition (18) can be written as (cf., e.g., [60]):

$$p(t)\Delta_q u(t) + (q-1) \sum_{s \in [t_0, t) \cap \mathbb{T}} sr(s)u(qs) = const. \tag{28}$$

Corollary 5 *If a given q-equation*

$$H(t, y(qt), \Delta_q y(t)) + (q-1) \sum_{s \in [t_0, t) \cap \mathbb{T}} sG(s, y(qs), \Delta_q y(s)) = const,$$

$q > 1$*, is to be the Euler–Lagrange equation of the variational problem*

$$\mathscr{I}[y] = (q-1) \sum_{t \in [t_0, t_1) \cap \mathbb{T}} tL(t, y(qt), \Delta_q y(t)),$$

$t_0, t_1 \in \overline{q^{\mathbb{Z}}}$*, then its equation of variation*

$$\begin{aligned} &H_y[u]_q(t)u(qt) + H_v[u]_q(t)\Delta_q u(t) \\ &\quad + (q-1) \sum_{s \in [t_0, t) \cap \mathbb{T}} s\left(G_y[u]_q(s)u(qs) + G_v[u]_q(s)\Delta_q u(s)\right) = 0 \end{aligned}$$

is self-adjoint, in the sense of Definition 15 with (18) *given by* (28).

Proof Choose $\mathbb{T} = \overline{q^{\mathbb{Z}}}$ in Theorem 17.

More information about Euler–Lagrange equations for q-variational problems may be found in [30, 48, 52] and references therein.

4.3 Discussion

On an arbitrary time scale $\mathbb{T}$, we can easily show equivalence between the integro-differential equation (20) and the second order differential equation (29) below (Proposition 1). However, when we consider equations of variations of them, we notice that it is not possible to prove an equivalence between them on an arbitrary time scale. The main reason of this impossibility, even in the discrete time scale $\mathbb{Z}$, is the absence of a general chain rule on an arbitrary time scale (see, e.g., Example 1.85 of [13]). However, on $\mathbb{T} = \mathbb{R}$ we can present this equivalence (Proposition 2).

Proposition 1 (See [26]) *The integro-differential equation* (20) *is equivalent to a second order delta differential equation*

$$W\left(t, y^{\sigma}(t), y^{\Delta}(t), y^{\Delta\Delta}(t)\right) = 0. \tag{29}$$

Let $\mathbb{T}$ be a time scale such that μ is delta differentiable. The equation of variation of a second order differential equation (29) is given by

$$W_z\langle u\rangle(t)u^{\Delta\Delta}(t) + W_v\langle u\rangle(t)u^{\Delta}(t) + W_y\langle u\rangle(t)u^{\sigma}(t) = 0. \tag{30}$$

On an arbitrary time scale it is impossible to prove the equivalence between the equation of variation (21) and (30). Indeed, after differentiating both sides of Eq. (21) and using the product rule given by Theorem 2, one has

$$\begin{aligned} H_y[u](t)u^{\sigma\Delta}(t) + H_y^{\Delta}[u](t)u^{\sigma\sigma}(t) + H_v[u](t)u^{\Delta\Delta}(t) + H_v^{\Delta}[u](t)u^{\Delta\sigma}(t) \\ + G_y[u](t)u^{\sigma}(t) + G_v[u](t)u^{\Delta}(t) = 0. \end{aligned} \tag{31}$$

The direct calculations

- $H_y[u](t)u^{\sigma\Delta}(t) = H_y[u](t)(u^{\Delta}(t) + \mu^{\Delta}(t)u^{\Delta}(t) + \mu^{\sigma}(t)u^{\Delta\Delta}(t))$,
- $H_y^{\Delta}[u](t)u^{\sigma\sigma}(t) = H_y^{\Delta}[u](t)(u^{\sigma}(t) + \mu^{\sigma}(t)u^{\Delta}(t) + \mu(t)\mu^{\sigma}(t)u^{\Delta\Delta}(t))$,
- $H_v^{\Delta}[u](t)u^{\Delta\sigma}(t) = H_v^{\Delta}[u](t)(u^{\Delta}(t) + \mu u^{\Delta\Delta}(t))$,

and the fourth item of Theorem 1, allow us to write Eq. (31) in form

$$\begin{aligned} u^{\Delta\Delta}(t)\left[\mu(t)H_y[u](t) + H_v[u](t)\right]^{\sigma} \\ + u^{\Delta}(t)\left[H_y[u](t) + (\mu(t)H_y[u](t))^{\Delta} + H_v^{\Delta}[u](t) + G_v[u](t)\right] \\ + u^{\sigma}(t)\left[H_y^{\Delta}[u](t) + G_y[u](t)\right] = 0. \end{aligned} \tag{32}$$

We are not able to prove that the coefficients of (32) are the same as in (30), respectively. This is due to the fact that we cannot find the partial derivatives of (29), that is, $W_z\langle u\rangle(t)$, $W_v\langle u\rangle(t)$ and $W_y\langle u\rangle(t)$, from Eq. (30) because of lack of a general chain rule in an arbitrary time scale [12]. The equivalence, however, is true for $\mathbb{T} = \mathbb{R}$. Operator $\langle\cdot\rangle$ has in this case the form $\langle y\rangle(t) = (t, y(t), y'(t), y''(t)) =: \langle y\rangle_{\mathbb{R}}(t)$.

Proposition 2 (See [26]) *The equation of variation*

$$H_y[u]_{\mathbb{R}}(t)u(t) + H_v[u]_{\mathbb{R}}(t)u'(t) + \int_{t_0}^{t} G_y[u]_{\mathbb{R}}(s)u(s) + G_v[u]_{\mathbb{R}}(s)u'(s)ds = 0$$

is equivalent to the second order differential equation

$$W_z\langle u\rangle_{\mathbb{R}}(t)u''(t) + W_v\langle u\rangle_{\mathbb{R}}(t)u'(t) + W_y\langle u\rangle_{\mathbb{R}}(t)u(t) = 0.$$

Proposition 2 allows us to obtain the classical result of [21, Theorem II] as a corollary of our Theorem 17. The absence of a chain rule on an arbitrary time scale (even for $\mathbb{T} = \mathbb{Z}$) implies that the classical approach [21] fails on time scales. This is the reason why we use here a completely different approach to the subject based on the integro-differential form. The case $\mathbb{T} = \mathbb{Z}$ was recently investigated in [17]. However, similarly to [21], the approach of [17] is based on the differential form and cannot be extended to general time scales.

5 The Delta-Nabla Calculus of Variations for Composition Functionals

The delta-nabla calculus of variations has been introduced in [44]. Here we investigate more general problems of the time-scale calculus of variations for a functional that is the composition of a certain scalar function with the delta and nabla integrals of a vector valued field. We begin by proving general Euler–Lagrange equations in integral form (Theorem 18). Then we consider cases when initial or terminal boundary conditions are not specified, obtaining corresponding transversality conditions (Theorems 19 and 20). Furthermore, we prove necessary optimality conditions for general isoperimetric problems given by the composition of delta-nabla integrals (Theorem 21). Finally, some illustrating examples are presented (Sect. 5.4).

5.1 The Euler–Lagrange Equations

Let us begin by defining the class of functions $C^1_{k,n}([a,b];\mathbb{R})$, which contains delta and nabla differentiable functions.

Definition 17 By $C^1_{k,n}([a,b];\mathbb{R})$, $k,n \in \mathbb{N}$, we denote the class of functions $y : [a,b] \to \mathbb{R}$ such that: if $k \neq 0$ and $n \neq 0$, then y^{Δ} is continuous on $[a,b]^{\kappa}_{\kappa}$ and y^{∇} is continuous on $[a,b]^{\kappa}_{\kappa}$, where $[a,b]^{\kappa}_{\kappa} := [a,b]^{\kappa} \cap [a,b]_{\kappa}$; if $n = 0$, then y^{Δ} is continuous on $[a,b]^{\kappa}$; if $k = 0$, then y^{∇} is continuous on $[a,b]_{\kappa}$.

Our aim is to find a function y which minimizes or maximizes the following variational problem:

$$\mathscr{L}[y] = H\left(\int_a^b f_1(t, y^{\sigma}(t), y^{\Delta}(t))\Delta t, \ldots, \int_a^b f_k(t, y^{\sigma}(t), y^{\Delta}(t))\Delta t, \right.$$
$$\left. \int_a^b f_{k+1}(t, y^{\rho}(t), y^{\nabla}(t))\nabla t, \ldots, \int_a^b f_{k+n}(t, y^{\rho}(t), y^{\nabla}(t))\nabla t\right), \tag{33}$$

$$(y(a) = y_a), \quad (y(b) = y_b). \tag{34}$$

The parentheses in (34), around the end-point conditions, means that those conditions may or may not occur (it is possible that one or both $y(a)$ and $y(b)$ are free). A function $y \in C^1_{k,n}$ is said to be admissible provided it satisfies the boundary conditions (34) (if any is given). For $k = 0$ problem (33)–(34) becomes a nabla problem (neither delta integral nor delta derivative is present); for $n = 0$ problem (33)–(34) reduces to a delta problem (neither nabla integral nor nabla derivative is present). For simplicity, we use the operators $[\cdot]$ and $\{\cdot\}$ defined by

$$[y](t) := (t, y^\sigma(t), y^\Delta(t)), \quad \{y\}(t) := (t, y^\rho(t), y^\nabla(t)).$$

We assume that:

1. the function $H : \mathbb{R}^{n+k} \to \mathbb{R}$ has continuous partial derivatives with respect to its arguments, which we denote by $H_i^{'}$, $i = 1, \ldots, n+k$;
2. functions $(t, y, v) \to f_i(t, y, v)$ from $[a, b] \times \mathbb{R}^2$ to $\mathbb{R}$, $i = 1, \ldots, n+k$, have partial continuous derivatives with respect to y and v uniformly in $t \in [a, b]$, which we denote by f_{iy} and f_{iv};
3. f_i, f_{iy}, f_{iv} are rd-continuous on $[a, b]^\kappa$, $i = 1, \ldots, k$, and ld-continuous on $[a, b]_\kappa$, $i = k+1, \ldots, k+n$, for all $y \in C^1_{k,n}$.

Definition 18 (*Cf.* [44]) We say that an admissible function $\hat{y} \in C^1_{k,n}([a, b]; \mathbb{R})$ is a local minimizer (respectively, local maximizer) to problem (33)–(34), if there exists $\delta > 0$ such that $\mathscr{L}[\hat{y}] \leq \mathscr{L}[y]$ (respectively, $\mathscr{L}[\hat{y}] \geq \mathscr{L}[y]$) for all admissible functions $y \in C^1_{k,n}([a, b]; \mathbb{R})$ satisfying the inequality $||y - \hat{y}||_{1,\infty} < \delta$, where

$$||y||_{1,\infty} := ||y^\sigma||_\infty + ||y^\Delta||_\infty + ||y^\rho||_\infty + ||y^\nabla||_\infty$$

with $||y||_\infty := \sup_{t \in [a,b]^\kappa_\kappa} |y(t)|$.

For brevity, in what follows we omit the argument of $H_i^{'}$. Precisely,

$$H_i^{'} := \frac{\partial H}{\partial \mathscr{F}_i}(\mathscr{F}_1(y), \ldots, \mathscr{F}_{k+n}(y)), \quad i = 1, \ldots, n+k,$$

where

$$\mathscr{F}_i(y) = \int\limits_a^b f_i[y](t)\Delta t, \text{ for } i = 1, \ldots, k,$$

$$\mathscr{F}_i(y) = \int\limits_a^b f_i\{y\}(t)\nabla t, \text{ for } i = k+1, \ldots, k+n.$$

Depending on the given boundary conditions, we can distinguish four different problems. The first one is the problem (P_{ab}), where the two boundary conditions are specified. To solve this problem we need an Euler–Lagrange necessary optimality condition, which is given by Theorem 18 below. Next two problems — denoted by (P_a) and (P_b) — occur when $y(a)$ is given and $y(b)$ is free (problem (P_a)) and when $y(a)$ is free and $y(b)$ is specified (problem (P_b)). To solve both of them we need an Euler–Lagrange equation and one proper transversality condition. The last problem — denoted by (P) — occurs when both boundary conditions are not present. To find a solution for such a problem we need to use an Euler–Lagrange equation and two transversality conditions (one at each time a and b).

Theorem 18 (The Euler–Lagrange equations in integral form) *If $\hat{y}$ is a local solution to problem* (33)–(34)*, then the Euler–Lagrange equations (in integral form)*

$$\sum_{i=1}^{k} H_i^{'} \cdot \left(f_{iv}[\hat{y}](\rho(t)) - \int_a^{\rho(t)} f_{iy}[\hat{y}](\tau)\Delta\tau \right) + \sum_{i=k+1}^{k+n} H_i^{'} \cdot \left(f_{iv}\{\hat{y}\}(t) - \int_a^{t} f_{iy}\{\hat{y}\}(\tau)\nabla\tau \right) = c, \quad t \in \mathbb{T}_\kappa, \tag{35}$$

and

$$\sum_{i=1}^{k} H_i^{'} \cdot \left(f_{iv}[\hat{y}](t) - \int_a^{t} f_{iy}[\hat{y}](\tau)\Delta\tau \right) + \sum_{i=k+1}^{k+n} H_i^{'} \cdot \left(f_{iv}\{\hat{y}\}(\sigma(t)) - \int_a^{\sigma(t)} f_{iy}\{\hat{y}\}(\tau)\nabla\tau \right) = c, \quad t \in \mathbb{T}^\kappa, \tag{36}$$

hold.

Proof See [25].

For regular time scales (Definition 3), the Euler–Lagrange equations (35) and (36) coincide; on a general time scale, they are different. Such a difference is illustrated in Example 8. For such purpose let us define ξ and χ by

$$\begin{aligned}\xi(t) &:= \sum_{i=1}^{k} H_i^{'} \cdot \left(f_{iv}[\hat{y}](t) - \int\limits_a^t f_{iy}[\hat{y}](\tau)\Delta\tau \right),\\ \chi(t) &:= \sum_{i=k+1}^{k+n} H_i^{'} \cdot \left(f_{iv}\{\hat{y}\}(t) - \int\limits_a^t f_{iy}\{\hat{y}\}(\tau)\nabla\tau \right).\end{aligned} \tag{37}$$

Example 8 Let us consider the irregular time scale $\mathbb{T} = \mathbb{P}_{1,1} = \bigcup\limits_{k=0}^{\infty} [2k, 2k+1]$. We show that for this time scale there is a difference between the Euler–Lagrange equations (35) and (36). The forward and backward jump operators are given by

$$\sigma(t) = \begin{cases} t, & t \in \bigcup\limits_{k=0}^{\infty} [2k, 2k+1), \\ t+1, & t \in \bigcup\limits_{k=0}^{\infty} \{2k+1\}, \end{cases} \qquad \rho(t) = \begin{cases} t, & t \in \bigcup\limits_{k=0}^{\infty} (2k, 2k+1], \\ t-1, & t \in \bigcup\limits_{k=1}^{\infty} \{2k\}, \\ 0, & t = 0. \end{cases}$$

For $t = 0$ and $t \in \bigcup\limits_{k=0}^{\infty} (2k, 2k+1)$, Eqs. (35) and (36) coincide. We can distinguish between them for $t \in \bigcup\limits_{k=0}^{\infty} \{2k+1\}$ and $t \in \bigcup\limits_{k=1}^{\infty} \{2k\}$. In what follows we use the notations (37). If $t \in \bigcup\limits_{k=0}^{\infty} \{2k+1\}$, then we obtain from (35) and (36) the Euler–Lagrange equations $\xi(t) + \chi(t) = c$ and $\xi(t) + \chi(t+1) = c$, respectively. If $t \in \bigcup\limits_{k=1}^{\infty} \{2k\}$, then the Euler–Lagrange equation (35) has the form $\xi(t-1) + \chi(t) = c$ while (36) takes the form $\xi(t) + \chi(t) = c$.

5.2 Natural Boundary Conditions

In this section we minimize or maximize the variational functional (33), but initial and/or terminal boundary condition $y(a)$ and/or $y(b)$ are not specified. In what follows we obtain corresponding transversality conditions.

Theorem 19 (Transversality condition at the initial time $t = a$) *Let $\mathbb{T}$ be a time scale for which $\rho(\sigma(a)) = a$. If $\hat{y}$ is a local extremizer to* (33) *with $y(a)$ not specified, then*

$$\sum_{i=1}^{k} H_i^{'} \cdot f_{iv}[\hat{y}](a) + \sum_{i=k+1}^{k+n} H_i^{'} \cdot \left(f_{iv}\{\hat{y}\}(\sigma(a)) - \int\limits_a^{\sigma(a)} f_{iy}\{\hat{y}\}(t)\nabla t \right) = 0$$

holds together with the Euler–Lagrange equations (35) *and* (36)*.*

Proof See [25].

Theorem 20 (Transversality condition at the terminal time $t = b$) *Let* $\mathbb{T}$ *be a time scale for which* $\sigma(\rho(b)) = b$*. If* $\hat{y}$ *is a local extremizer to* (33) *with* $y(b)$ *not specified, then*

$$\sum_{i=1}^{k} H_i^{'} \cdot \left(f_{iv}[\hat{y}](\rho(b)) + \int_{\rho(b)}^{b} f_{iy}[\hat{y}](t)\Delta t \right) + \sum_{i=k+1}^{k+n} H_i^{'} \cdot f_{iv}\{\hat{y}\}(b) = 0$$

holds together with the Euler–Lagrange equations (35) *and* (36)*.*

Proof See [25].

Several interesting results can be immediately obtained from Theorems 18–20. An example of such results is given by Corollary 6.

Corollary 6 *If* $\hat{y}$ *is a solution to the problem*

$$\mathscr{L}[y] = \frac{\int_a^b f_1(t, y^\sigma(t), y^\Delta(t))\Delta t}{\int_a^b f_2(t, y^\rho(t), y^\nabla(t))\nabla t} \longrightarrow \text{extr},$$

$$(y(a) = y_a), \quad (y(b) = y_b),$$

then the Euler–Lagrange equations

$$\frac{1}{\mathscr{F}_2}\left(f_{1v}[\hat{y}](\rho(t)) - \int_a^{\rho(t)} f_{1y}[\hat{y}](\tau)\Delta\tau \right) - \frac{\mathscr{F}_1}{\mathscr{F}_2^2}\left(f_{2v}\{\hat{y}\}(t) - \int_a^{t} f_{2y}\{\hat{y}\}(\tau)\nabla\tau \right) = c,$$

$t \in \mathbb{T}_\kappa$*, and*

$$\frac{1}{\mathscr{F}_2}\left(f_{1v}[\hat{y}](t) - \int_a^{t} f_{1y}[\hat{y}](\tau)\Delta\tau \right) - \frac{\mathscr{F}_1}{\mathscr{F}_2^2}\left(f_{2v}\{\hat{y}\}(\sigma(t)) - \int_a^{\sigma(t)} f_{2y}\{\hat{y}\}(\tau)\nabla\tau \right) = c,$$

$t \in \mathbb{T}^\kappa$*, hold, where*

$$\mathscr{F}_1 := \int_a^b f_1(t, \hat{y}^\sigma(t), \hat{y}^\Delta(t))\Delta t \quad \textit{and} \quad \mathscr{F}_2 := \int_a^b f_2(t, \hat{y}^\rho(t), \hat{y}^\nabla(t))\nabla t.$$

Moreover, if $y(a)$ is free and $\rho(\sigma(a)) = a$, then

$$\frac{1}{\mathscr{F}_2} f_{1v}[\hat{y}](a) - \frac{\mathscr{F}_1}{\mathscr{F}_2^2}\left(f_{2v}\{\hat{y}\}(\sigma(a)) - \int_a^{\sigma(a)} f_{2y}\{\hat{y}\}(t)\nabla t \right) = 0;$$

if $y(b)$ is free and $\sigma(\rho(b)) = b$, then

$$\frac{1}{\mathscr{F}_2}\left(f_{1v}[\hat{y}](\rho(b)) + \int_{\rho(b)}^{b} f_{1y}[\hat{y}](t)\Delta t \right) - \frac{\mathscr{F}_1}{\mathscr{F}_2^2} f_{2v}\{\hat{y}\}(b) = 0.$$

5.3 Isoperimetric Problems

Let us now consider the general delta–nabla composition isoperimetric problem on time scales subject to boundary conditions. The problem consists of extremizing

$$\begin{aligned}\mathscr{L}[y] = H \Bigg(&\int_a^b f_1(t, y^\sigma(t), y^\Delta(t))\Delta t, \ldots, \int_a^b f_k(t, y^\sigma(t), y^\Delta(t))\Delta t, \\ &\int_a^b f_{k+1}(t, y^\rho(t), y^\nabla(t))\nabla t, \ldots, \int_a^b f_{k+n}(t, y^\rho(t), y^\nabla(t))\nabla t \Bigg)\end{aligned} \tag{38}$$

in the class of functions $y \in C^1_{k+m,n+p}$ satisfying given boundary conditions

$$y(a) = y_a, \quad y(b) = y_b, \tag{39}$$

and a generalized isoperimetric constraint

$$\begin{aligned}\mathscr{K}[y] = P \Bigg(&\int_a^b g_1(t, y^\sigma(t), y^\Delta(t))\Delta t, \ldots, \int_a^b g_m(t, y^\sigma(t), y^\Delta(t))\Delta t, \\ &\int_a^b g_{m+1}(t, y^\rho(t), y^\nabla(t))\nabla t, \ldots, \int_a^b g_{m+p}(t, y^\rho(t), y^\nabla(t))\nabla t \Bigg) = d,\end{aligned} \tag{40}$$

where $y_a, y_b, d \in \mathbb{R}$. We assume that:

1. the functions $H : \mathbb{R}^{n+k} \to \mathbb{R}$ and $P : \mathbb{R}^{m+p} \to \mathbb{R}$ have continuous partial derivatives with respect to all their arguments, which we denote by $H_i^{'}, i=1, \ldots, n+k$, and $P_i^{'}, i = 1, \ldots, m+p$;
2. functions $(t, y, v) \to f_i(t, y, v)$, $i = 1, \ldots, n+k$, and $(t, y, v) \to g_j(t, y, v)$, $j = 1, \ldots, m+p$, from $[a, b] \times \mathbb{R}^2$ to $\mathbb{R}$, have partial continuous derivatives with respect to y and v uniformly in $t \in [a, b]$, which we denote by f_{iy}, f_{iv}, and g_{jy}, g_{jv};
3. for all $y \in C^1_{k+m,n+p}$, f_i, f_{iy}, f_{iv} and g_j, g_{jy}, g_{jv} are rd-continuous in $t \in [a, b]^\kappa$, $i = 1, \ldots, k$, $j = 1, \ldots, m$, and ld-continuous in $t \in [a, b]_\kappa$, $i = k+1, \ldots, k+n$, $j = m+1, \ldots, m+p$.

A function $y \in C^1_{k+m,n+p}$ is said to be admissible provided it satisfies the boundary conditions (39) and the isoperimetric constraint (40). For brevity, we omit the argument of $P_i^{'}$: $P_i^{'} := \frac{\partial P}{\partial \mathscr{G}_i}(\mathscr{G}_1(\hat{y}), \ldots, \mathscr{G}_{m+p}(\hat{y}))$ for $i = 1, \ldots, m+p$, with

$$\mathscr{G}_i(\hat{y}) = \int_a^b g_i(t, \hat{y}^\sigma(t), \hat{y}^\Delta(t))\Delta t, \quad i = 1, \ldots, m,$$

and

$$\mathscr{G}_i(\hat{y}) = \int_a^b g_i(t, \hat{y}^\rho(t), \hat{y}^\nabla(t))\nabla t, \quad i = m+1, \ldots, m+p.$$

Definition 19 We say that an admissible function $\hat{y}$ is a local minimizer (respectively, a local maximizer) to the isoperimetric problem (38)–(40), if there exists a $\delta > 0$ such that $\mathscr{L}[\hat{y}] \leqslant \mathscr{L}[y]$ (respectively, $\mathscr{L}[\hat{y}] \geqslant \mathscr{L}[y]$) for all admissible functions $y \in C^1_{k+m,n+p}$ satisfying the inequality $||y - \hat{y}||_{1,\infty} < \delta$.

Let us define u and w by

$$\begin{aligned} u(t) &:= \sum_{i=1}^{m} P_i^{'} \cdot \left(g_{iv}[\hat{y}](t) - \int_a^t g_{iy}[\hat{y}](\tau)\Delta\tau \right), \\ w(t) &:= \sum_{i=m+1}^{m+p} P_i^{'} \cdot \left(g_{iv}\{\hat{y}\}(t) - \int_a^t g_{iy}\{\hat{y}\}(\tau)\nabla\tau \right). \end{aligned} \tag{41}$$

Definition 20 An admissible function $\hat{y}$ is said to be an extremal for $\mathscr{K}$ if $u(t) + w(\sigma(t)) = const$ and $u(\rho(t)) + w(t) = const$ for all $t \in [a, b]^\kappa_\kappa$. An extremizer (i.e., a local minimizer or a local maximizer) to problem (38)–(40) that is not an extremal for $\mathscr{K}$ is said to be a normal extremizer; otherwise (i.e., if it is an extremal for $\mathscr{K}$), the extremizer is said to be abnormal.

Theorem 21 (Optimality condition to the isoperimetric problem (38)–(40)) *Let χ and ξ be given as in (37), and u and w be given as in (41). If $\hat{y}$ is a normal extremizer to the isoperimetric problem (38)–(40), then there exists a real number λ such that*

1. $\xi^{\rho}(t) + \chi(t) - \lambda\left(u^{\rho}(t) + w(t)\right) = const;$
2. $\xi(t) + \chi^{\sigma}(t) - \lambda\left(u^{\rho}(t) + w(t)\right) = const;$
3. $\xi^{\rho}(t) + \chi(t) - \lambda\left(u(t) + w^{\sigma}(t)\right) = const;$
4. $\xi(t) + \chi^{\sigma}(t) - \lambda\left(u(t) + w^{\sigma}(t)\right) = const;$

for all $t \in [a, b]^{\kappa}_{\kappa}$.

Proof See proof of Theorem 3.9 in [25].

5.4 Illustrative Examples

In this section we consider three examples which illustrate the results of Theorems 18 and 21. We begin with a nonautonomous problem.

Example 9 Consider the problem

$$\mathscr{L}[y] = \frac{\int\limits_0^1 t y^{\Delta}(t)\Delta t}{\int\limits_0^1 (y^{\nabla}(t))^2 \nabla t} \longrightarrow \min, \qquad y(0) = 0, \quad y(1) = 1. \tag{42}$$

If y is a local minimizer to problem (42), then the Euler–Lagrange equations of Corollary 6 must hold, i.e.,

$$\frac{1}{\mathscr{F}_2}\rho(t) - 2\frac{\mathscr{F}_1}{\mathscr{F}_2^2} y^{\nabla}(t) = c, \quad t \in \mathbb{T}_{\kappa},$$

and

$$\frac{1}{\mathscr{F}_2} t - 2\frac{\mathscr{F}_1}{\mathscr{F}_2^2} y^{\nabla}(\sigma(t)) = c, \quad t \in \mathbb{T}^{\kappa},$$

where $\mathscr{F}_1 := \mathscr{F}_1(y) = \int\limits_0^1 t y^{\Delta}(t)\Delta t$ and $\mathscr{F}_2 := \mathscr{F}_2(y) = \int\limits_0^1 (y^{\nabla}(t))^2 \nabla t$. Let us consider the second equation. Using (4) of Theorem 9, it can be written as

$$\frac{1}{\mathscr{F}_2} t - 2\frac{\mathscr{F}_1}{\mathscr{F}_2^2} y^{\Delta}(t) = c, \quad t \in \mathbb{T}^{\kappa}. \tag{43}$$

Solving (43) subject to the boundary conditions $y(0) = 0$ and $y(1) = 1$ gives

$$y(t) = \frac{1}{2Q}\int_0^t \tau\Delta\tau - t\left(\frac{1}{2Q}\int_0^1 \tau\Delta\tau - 1\right), \quad t \in \mathbb{T}^{\kappa}, \tag{44}$$

where $Q := \frac{\mathscr{F}_1}{\mathscr{F}_2}$. Therefore, the solution depends on the time scale. Let us consider two examples: $\mathbb{T} = \mathbb{R}$ and $\mathbb{T} = \left\{0, \frac{1}{2}, 1\right\}$. On $\mathbb{T} = \mathbb{R}$, from (44) we obtain

$$y(t) = \frac{1}{4Q}t^2 + \frac{4Q-1}{4Q}t, \qquad y^{\Delta}(t) = y^{\nabla}(t) = y'(t) = \frac{1}{2Q}t + \frac{4Q-1}{4Q} \tag{45}$$

as solution of (43). Substituting (45) into $\mathscr{F}_1$ and $\mathscr{F}_2$ gives $\mathscr{F}_1 = \frac{12Q+1}{24Q}$ and $\mathscr{F}_2 = \frac{48Q^2+1}{48Q^2}$, that is,

$$Q = \frac{2Q(12Q+1)}{48Q^2+1}. \tag{46}$$

Solving Eq. (46) we get $Q \in \left\{\frac{3-2\sqrt{3}}{12}, \frac{3+2\sqrt{3}}{12}\right\}$. Because (42) is a minimizing problem, we select $Q = \frac{3-2\sqrt{3}}{12}$ and we get the extremal

$$y(t) = -(3+2\sqrt{3})t^2 + (4+2\sqrt{3})t. \tag{47}$$

If $\mathbb{T} = \left\{0, \frac{1}{2}, 1\right\}$, then from (44) we obtain $y(t) = \frac{1}{8Q}\sum_{k=0}^{2t-1} k + \frac{8Q-1}{8Q}t$, that is,

$$y(t) = \begin{cases} 0, & \text{if } t = 0, \\ \frac{8Q-1}{16Q}, & \text{if } t = \frac{1}{2}, \\ 1, & \text{if } t = 1. \end{cases}$$

Direct calculations show that

$$\begin{aligned} y^{\Delta}(0) &= \frac{y(\frac{1}{2}) - y(0)}{\frac{1}{2}} = \frac{8Q-1}{8Q}, \quad y^{\Delta}\left(\frac{1}{2}\right) = \frac{y(1) - y(\frac{1}{2})}{\frac{1}{2}} = \frac{8Q+1}{8Q}, \\ y^{\nabla}\left(\frac{1}{2}\right) &= \frac{y(\frac{1}{2}) - y(0)}{\frac{1}{2}} = \frac{8Q-1}{8Q}, \quad y^{\nabla}(1) = \frac{y(1) - y(\frac{1}{2})}{\frac{1}{2}} = \frac{8Q+1}{8Q}. \end{aligned} \tag{48}$$

Substituting (48) into the integrals $\mathscr{F}_1$ and $\mathscr{F}_2$ gives

$$\mathscr{F}_1 = \frac{8Q+1}{32Q}, \quad \mathscr{F}_2 = \frac{64Q^2+1}{64Q^2}, \quad Q = \frac{\mathscr{F}_1}{\mathscr{F}_2} = \frac{2Q(8Q+1)}{64Q^2+1}.$$

Thus, we obtain the equation $64Q^2 - 16Q - 1 = 0$. The solutions to this equation are: $Q \in \left\{\frac{1-\sqrt{2}}{8}, \frac{1+\sqrt{2}}{8}\right\}$. We are interested in the minimum value Q, so we select $Q = \frac{1+\sqrt{2}}{8}$ to get the extremal

$$y(t) = \begin{cases} 0, & \text{if } t = 0, \\ 1 - \frac{\sqrt{2}}{2}, & \text{if } t = \frac{1}{2}, \\ 1, & \text{if } t = 1. \end{cases} \tag{49}$$

Note that the extremals (47) and (49) are different: for (47) one has $x(1/2) = \frac{5}{4} + \frac{\sqrt{3}}{2}$.

In the previous example, the variational functional is given by the ratio of a delta and a nabla integral. Now we discuss a variational problem where the composition is expressed by the product of three time-scale integrals.

Example 10 Consider the problem

$$\begin{gathered} \mathscr{L}[y] = \left(\int_0^3 t y^{\Delta}(t) \Delta t\right)\left(\int_0^3 y^{\Delta}(t)\,(1+t)\,\Delta t\right)\left(\int_0^3 \left[\left(y^{\nabla}(t)\right)^2 + t\right] \nabla t\right) \longrightarrow \min, \\ y(0) = 0, \quad y(3) = 3. \end{gathered} \tag{50}$$

If y is a local minimizer to problem (50), then the Euler–Lagrange equations must hold, and we can write that

$$(\mathscr{F}_1\mathscr{F}_3 + \mathscr{F}_2\mathscr{F}_3)\, t + \mathscr{F}_1\mathscr{F}_3 + 2\mathscr{F}_1\mathscr{F}_2 y^{\nabla}(\sigma(t)) = c, \quad t \in \mathbb{T}^{\kappa}, \tag{51}$$

where c is a constant, $\mathscr{F}_1 := \mathscr{F}_1(y) = \int_0^3 t y^{\Delta}(t) \Delta t$, $\mathscr{F}_2 := \mathscr{F}_2(y) = \int_0^3 y^{\Delta}(t)$ $(1+t)\,\Delta t$, and $\mathscr{F}_3 := \mathscr{F}_3(y) = \int_0^3 \left[\left(y^{\nabla}(t)\right)^2 + t\right] \nabla t$. Using relation (4), we can write (51) as

$$(\mathscr{F}_1\mathscr{F}_3 + \mathscr{F}_2\mathscr{F}_3)\, t + \mathscr{F}_1\mathscr{F}_3 + 2\mathscr{F}_1\mathscr{F}_2 y^{\Delta}(t) = c, \quad t \in \mathbb{T}^{\kappa}. \tag{52}$$

Using the boundary conditions $y(0) = 0$ and $y(3) = 3$, from (52) we get that

$$y(t) = \left(1 + \frac{Q}{3}\int_0^3 \tau \Delta\tau\right) t - Q\int_0^t \tau \Delta\tau, \quad t \in \mathbb{T}^{\kappa}, \tag{53}$$

where $Q = \frac{\mathscr{F}_1\mathscr{F}_3 + \mathscr{F}_2\mathscr{F}_3}{2\mathscr{F}_1\mathscr{F}_2}$. Therefore, the solution depends on the time scale. Let us consider $\mathbb{T} = \mathbb{R}$ and $\mathbb{T} = \left\{0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3\right\}$. On $\mathbb{T} = \mathbb{R}$, expression (53) gives

$$y(t)=\left(\frac{2+3Q}{2}\right)t-\frac{Q}{2}t^2,\quad y^{\Delta}(t)=y^{\nabla}(t)=y'(t)=\frac{2+3Q}{2}-Qt \tag{54}$$

as solution of (52). Substituting (54) into $\mathscr{F}_1$, $\mathscr{F}_2$ and $\mathscr{F}_3$ gives:

$$\mathscr{F}_1=\frac{18-9Q}{4},\quad \mathscr{F}_2=\frac{30-9Q}{4},\quad \mathscr{F}_3=\frac{9Q^2+30}{4}.$$

Solving equation $9Q^3-36Q^2+45Q-40=0$, one finds the solution

$$Q=\frac{1}{27}\left[36+\sqrt[3]{24786-729\sqrt{1155}}+9\sqrt[3]{34+\sqrt{1155}}\right]\approx 2,7755$$

and the extremal $y(t)=5,16325t-1,38775t^2$.

Let us consider now the time scale $\mathbb{T}=\left\{0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},3\right\}$. From (53), we obtain

$$y(t)=\left(\frac{4+5Q}{4}\right)t-\frac{Q}{4}\sum_{k=0}^{2t-1}k=\begin{cases}0, & \text{if } t=0,\\ \frac{4+5Q}{8}, & \text{if } t=\frac{1}{2},\\ 1+Q, & \text{if } t=1,\\ \frac{12+9Q}{8}, & \text{if } t=\frac{3}{2},\\ 2+Q, & \text{if } t=2,\\ \frac{20+5Q}{8}, & \text{if } t=\frac{5}{2},\\ 3, & \text{if } t=3\end{cases} \tag{55}$$

as solution of (52). Substituting (55) into $\mathscr{F}_1$, $\mathscr{F}_2$ and $\mathscr{F}_3$, yields

$$\mathscr{F}_1=\frac{60-35Q}{16},\quad \mathscr{F}_2=\frac{108-35Q}{16},\quad \mathscr{F}_3=\frac{35Q^2+132}{16}.$$

Solving equation $245Q^3-882Q^2+1110-\frac{5544}{5}=0$, we get $Q\approx 2,5139$ and the extremal

$$y(t)=\begin{cases}0, & \text{if } t=0,\\ 2,0711875, & \text{if } t=\frac{1}{2},\\ 3,5139, & \text{if } t=1,\\ 4,3281375, & \text{if } t=\frac{3}{2},\\ 4,5139, & \text{if } t=2,\\ 4,0711875, & \text{if } t=\frac{5}{2},\\ 3, & \text{if } t=3\end{cases} \tag{56}$$

for problem (50) on $\mathbb{T}=\left\{0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},3\right\}$.

In order to illustrate the difference between composition of mixed delta-nabla integrals and pure delta or nabla situations, we consider now two variants of problem (50): (i) the first consisting of delta operators only:

$$\mathscr{L}[y]=\left(\int_0^3 ty^{\Delta}(t)\Delta t\right)\left(\int_0^3 y^{\Delta}(t)\,(1+t)\,\Delta t\right)\left(\int_0^3\left[\left(y^{\Delta}(t)\right)^2+t\right]\Delta t\right)\longrightarrow \min; \tag{57}$$

(ii) the second of nabla operators only:

$$\mathscr{L}[y]=\left(\int_0^3 ty^{\nabla}(t)\nabla t\right)\left(\int_0^3 y^{\nabla}(t)\,(1+t)\,\nabla t\right)\left(\int_0^3\left[\left(y^{\nabla}(t)\right)^2+t\right]\nabla t\right)\longrightarrow \min\,. \tag{58}$$

Both problems (i) and (ii) are subject to the same boundary conditions as in (50):

$$y(0)=0, \quad y(3)=3. \tag{59}$$

All three problems (50), (57) and (59), and (58), (59), coincide in $\mathbb{R}$. Consider, as before, the time scale $\mathbb{T}=\left\{0,\frac{1}{2},1,\frac{3}{2},2,\frac{5}{2},3\right\}$. Recall that problem (50) has extremal (56). (i) Now, let us consider the delta problem (57) and (59). We obtain

$$\mathscr{F}_1=\frac{60-35Q}{16}, \quad \mathscr{F}_2=\frac{108-35Q}{16}, \quad \mathscr{F}_3=\frac{35Q^2+108}{16}$$

and the equation $245Q^3-882Q^2+1026-\frac{5436}{5}=0$. Its numerical solution $Q\approx 2,5216$ entails the extremal

$$y(t)=\begin{cases}0, & \text{if } t=0,\\ 2,076, & \text{if } t=\frac{1}{2},\\ 3,5216, & \text{if } t=1,\\ 4,3368, & \text{if } t=\frac{3}{2},\\ 4,5216, & \text{if } t=2,\\ 4,076, & \text{if } t=\frac{5}{2},\\ 3, & \text{if } t=3.\end{cases}$$

(ii) In the latter nabla problem (58), (59) we have

$$\mathscr{F}_1=\frac{84-35Q}{16}, \quad \mathscr{F}_2=\frac{132-35Q}{16}, \quad \mathscr{F}_3=\frac{35Q^2+132}{16}$$

and the equation $175Q^3-810Q^2+1122-\frac{7128}{7}=0$. Using its numerical solution $Q\approx 3,1097$ we get the extremal

$$y(t)=\begin{cases}0, & \text{if } t=0,\\ 2,4942, & \text{if } t=\frac{1}{2},\\ 4,1907, & \text{if } t=1,\\ 5,0895, & \text{if } t=\frac{3}{2},\\ 5,1907, & \text{if } t=2,\\ 4,4942, & \text{if } t=\frac{5}{2},\\ 3, & \text{if } t=3.\end{cases}$$

Finally, we apply the results of Sect. 5.3 to an isoperimetric problem.

Example 11 Let us consider the problem of extremizing

$$\mathscr{L}[y]=\frac{\int\limits_0^1 (y^{\Delta}(t))^2\Delta t}{\int\limits_0^1 t y^{\nabla}(t)\nabla t}$$

subject to the boundary conditions $y(0)=0$ and $y(1)=1$ and the isoperimetric constraint

$$\mathscr{K}[y]=\int\limits_0^1 t y^{\nabla}(t)\nabla t=1.$$

Applying Theorem 21, we get the nabla differential equation

$$\frac{2}{\mathscr{F}_2}y^{\nabla}(t)-\left(\lambda+\frac{\mathscr{F}_1}{(\mathscr{F}_2)^2}\right)t=c,\quad t\in\mathbb{T}^{\kappa}_{\kappa}. \tag{60}$$

Solving this equation, we obtain

$$y(t)=\left(1-Q\int\limits_0^1 \tau\nabla\tau\right)t+Q\int\limits_0^t \tau\nabla\tau, \tag{61}$$

where $Q=\frac{\mathscr{F}_2}{2}\left(\frac{\mathscr{F}_1}{(\mathscr{F}_2)^2}+\lambda\right)$. Therefore, the solution of equation (60) depends on the time scale. Let us consider $\mathbb{T}=\mathbb{R}$ and $\mathbb{T}=\left\{0,\frac{1}{2},1\right\}$.

On $\mathbb{T}=\mathbb{R}$, from (61) we obtain that $y(t)=\frac{2-Q}{2}t+\frac{Q}{2}t^2$. Substituting this expression for y into the integrals $\mathscr{F}_1$ and $\mathscr{F}_2$, gives $\mathscr{F}_1=\frac{Q^2+12}{12}$ and $\mathscr{F}_2=\frac{Q+6}{12}$. Using the given isoperimetric constraint, we obtain $Q=6$, $\lambda=8$, and $y(t)=3t^2-2t$. Let us consider now the time scale $\mathbb{T}=\left\{0,\frac{1}{2},1\right\}$. From (61), we have

$$y(t) = \frac{4-3Q}{4}t + Q\sum_{k=1}^{2t}\frac{k}{4} = \begin{cases} 0, & \text{if } t = 0,\\ \frac{4-Q}{8}, & \text{if } t = \frac{1}{2},\\ 1, & \text{if } t = 1. \end{cases}$$

Simple calculations show that

$$\mathscr{F}_1 = \sum_{k=0}^{1}\frac{1}{2}\left(y^{\Delta}\left(\frac{k}{2}\right)\right)^2 = \frac{1}{2}\left(y^{\Delta}(0)\right)^2 + \frac{1}{2}\left(y^{\Delta}\left(\frac{1}{2}\right)\right)^2 = \frac{Q^2+16}{16},$$

$$\mathscr{F}_2 = \sum_{k=1}^{2}\frac{1}{4}ky^{\nabla}\left(\frac{k}{2}\right) = \frac{1}{4}y^{\nabla}\left(\frac{1}{2}\right) + \frac{1}{2}y^{\nabla}(1) = \frac{Q+12}{16}$$

and $\mathscr{K}(y) = \frac{Q+12}{16} = 1$. Therefore, $Q = 4$, $\lambda = 6$, and we have the extremal

$$y(t) = \begin{cases} 0, & \text{if } t \in \left\{0, \frac{1}{2}\right\},\\ 1, & \text{if } t = 1. \end{cases}$$

6 Conclusions

In this survey we collected some of our recent research on direct and inverse problems of the calculus of variations on arbitrary time scales. For infinity horizon variational problems on time scales we refer the reader to [24, 53]. We started by studying inverse problems of the calculus of variations, which have not been studied before in the time-scale framework. First we derived a general form of a variational functional which attains a local minimum at a given function y_0 under Euler–Lagrange and strengthened Legendre conditions (Theorem 16). Next we considered a new approach to the inverse problem of the calculus of variations by using an integral perspective instead of the classical differential point of view. In order to solve the problem, we introduced new definitions: (i) self-adjointness of an integro-differential equation, and (ii) equation of variation. We obtained a necessary condition for an integro-differential equation to be an Euler–Lagrange equation on an arbitrary time scale $\mathbb{T}$ (Theorem 17). It remains open the question of sufficiency. Finally, we developed the direct calculus of variations by considering functionals that are a composition of a certain scalar function with delta and nabla integrals of a vector valued field. For such problems we obtained delta-nabla Euler–Lagrange equations in integral form (Theorem 18), transversality conditions (Theorems 19 and 20) and necessary optimality conditions for isoperimetric problems (Theorem 21). To consider such general mixed delta-nabla variational problems on unbounded time scales (infinite horizon problems) remains also an open direction of research. Another interesting open research direction consists to study delta-nabla inverse problems of calculus of variations for composition functionals and their conservation laws [61].

Acknowledgements This work was partially supported by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (CIDMA), and *The Portuguese Foundation for Science and Technology* (FCT), within project UID/MAT/04106/2013. The authors are grateful to two anonymous referees for their valuable comments and suggestions.

References

1. Ahlbrandt, C.D., Morian, C.: Partial differential equations on time scales. J. Comput. Appl. Math. **141**(1–2), 35–55 (2002)
2. Albu, I.D., Opriş, D.: Helmholtz type condition for mechanical integrators. Novi Sad J. Math. **29**(3), 11–21 (1999)
3. Almeida, R., Torres, D.F.M.: Isoperimetric problems on time scales with nabla derivatives. J. Vib. Control **15**(6), 951–958 (2009)
4. Almeida, R., Pooseh, S., Torres, D.F.M.: Computational Methods in the Fractional Calculus of Variations. Imperial College Press, London (2015)
5. Atici, F.M., Guseinov, G.Sh.: On Green's functions and positive solutions for boundary value problems on time scales. J. Comput. Appl. Math. **141**(1–2), 75–99 (2002)
6. Atici, F.M., McMahan, C.S.: A comparison in the theory of calculus of variations on time scales with an application to the Ramsey model. Nonlinear Dyn. Syst. Theory **9**(1), 1–10 (2009)
7. Atici, F.M., Uysal, F.: A production-inventory model of HMMS on time scales. Appl. Math. Lett. **21**(3), 236–243 (2008)
8. Atici, F.M., Biles, D.C., Lebedinsky, A.: An application of time scales to economics. Math. Comput. Model. **43**(7–8), 718–726 (2006)
9. Bartosiewicz, Z., Kotta, Ü., Pawłuszewicz, E., Wyrwas, M.: Control systems on regular time scales and their differential rings. Math. Control Signals Syst. **22**(3), 185–201 (2011)
10. Bastos, N.R.O., Ferreira, R.A.C., Torres, D.F.M.: Discrete-time fractional variational problems. Signal Process. **91**(3), 513–524 (2011)
11. Bohner, M.: Calculus of variations on time scales. Dyn. Syst. Appl. **13**(3–4), 339–349 (2004)
12. Bohner, M., Guseinov, G.Sh.: Partial differentiation on time scales. Dyn. Syst. Appl. **13**(3–4), 351–379 (2004)
13. Bohner, M., Peterson, A.: Dynamic Equations on Time Scales. Birkhäuser Boston, Boston (2001)
14. Bohner, M., Peterson, A.: Advances in Dynamic Equations on Time Scales. Birkhäuser Boston, Boston (2003)
15. Bohner, M., Guseinov, G., Peterson, A.: Introduction to the time scales calculus. In: Advances in Dynamic Equations on Time Scales, pp. 1–15. Birkhäuser Boston, Boston (2003)
16. Bohner, M.J., Ferreira, R.A.C., Torres, D.F.M.: Integral inequalities and their applications to the calculus of variations on time scales. Math. Inequal. Appl. **13**(3), 511–522 (2010)
17. Bourdin, L., Cresson, J.: Helmholtz's inverse problem of the discrete calculus of variations. J. Differ. Equ. Appl. **19**(9), 1417–1436 (2013)
18. Caputo, M.C.: Time scales: from nabla calculus to delta calculus and vice versa via duality. Int. J. Differ. Equ. **5**(1), 25–40 (2010)
19. Crăciun, D., Opriş, D.: The Helmholtz conditions for the difference equations systems. Balkan J. Geom. Appl. **1**(2), 21–30 (1996)
20. Darboux, G.: Leçons sur la Théorie Générale des Surfaces, Paris, (1894)
21. Davis, D.R.: The inverse problem of the calculus of variations in higher space. Trans. Amer. Math. Soc. **30**(4), 710–736 (1928)
22. Douglas, J.: Solution of the inverse problem of the calculus of variations. Trans. Amer. Math. Soc. **50**, 71–128 (1941)
23. Dryl, M.: Calculus of variations on time scales and applications to economics. Ph.D. Thesis, University of Aveiro (2014)

24. Dryl, M., Torres, D.F.M.: Necessary optimality conditions for infinite horizon variational problems on time scales. Numer. Algebra Control Optim. **3**(1), 145–160 (2013)
25. Dryl, M., Torres, D.F.M.: The delta-nabla calculus of variations for composition functionals on time scales. Int. J. Differ. Equ. **8**(1), 27–47 (2013)
26. Dryl, M., Torres, D.F.M.: Necessary condition for an Euler-Lagrange equation on time scales. Abstr. Appl. Anal. **7**, Art. ID 631281 (2014)
27. Dryl, M., Malinowska, A.B., Torres, D.F.M.: A time-scale variational approach to inflation, unemployment and social loss. Control Cybernet. **42**(2), 399–418 (2013)
28. Dryl, M., Malinowska, A.B., Torres, D.F.M.: An inverse problem of the calculus of variations on arbitrary time scales. Int. J. Differ. Equ. **9**(1), 53–66 (2014)
29. Ernst, T.: The different tongues of q-calculus. Proc. Est. Acad. Sci. **57**(2), 81–99 (2008)
30. Ferreira, R.A.C., Torres, D.F.M.: Necessary optimality conditions for the calculus of variations on time scales (2007). arXiv:0704.0656 [math.OC]
31. Ferreira, R.A.C., Torres, D.F.M.: Remarks on the calculus of variations on time scales. Int. J. Ecol. Econ. Stat. **9**(F07), 65–73 (2007)
32. Ferreira, R.A.C., Torres, D.F.M.: Isoperimetric problems of the calculus of variations on time scales. In: Nonlinear analysis and optimization II. Optimization. 123–131, Contemp. Math., 514, Amer. Math. Soc., Providence, RI (2010)
33. Girejko, E., Torres, D.F.M.: The existence of solutions for dynamic inclusions on time scales via duality. Appl. Math. Lett. **25**(11), 1632–1637 (2012)
34. Girejko, E., Malinowska, A.B., Torres, D.F.M.: The contingent epiderivative and the calculus of variations on time scales. Optimization **61**(3), 251–264 (2012)
35. Helmholtz, H. von: Ueber die physikalische Bedeutung des Prinicips der kleinsten Wirkung. J. Reine Angew. Math. **100**, 137–166 (1887)
36. Hilger, S.: Ein maßkettenkalkül mit anwendung auf zentrumsmannigfaltigkeiten. Ph.D. Thesis, Universität Würzburg (1988)
37. Hilger, S.: Analysis on measure chains–a unified approach to continuous and discrete calculus. Results Math. **18**(1–2), 18–56 (1990)
38. Hilger, S.: Differential and difference calculus–unified!. Nonlinear Anal. **30**(5), 2683–2694 (1997)
39. Hilscher, R., Zeidan, V.: Calculus of variations on time scales: weak local piecewise C^1_{rd} solutions with variable endpoints. J. Math. Anal. Appl. **289**(1), 143–166 (2004)
40. Hirsch, A.: Ueber eine charakteristische Eigenschaft der Differentialgleichungen der Variationsrechnung. Math. Ann. **49**(1), 49–72 (1897)
41. Hydon, P.E., Mansfield, E.L.: A variational complex for difference equations. Found. Comput. Math. **4**(2), 187–217 (2004)
42. Kac, V., Cheung, P.: Quantum Calculus. Universitext, Springer, New York (2002)
43. Lakshmikantham, V., Sivasundaram, S., Kaymakcalan, B.: Dynamic Systems on Measure Chains. Kluwer Academic Publishers, Dordrecht (1996)
44. Malinowska, A.B., Torres, D.F.M.: The delta-nabla calculus of variations. Fasc. Math. **44**, 75–83 (2010)
45. Malinowska, A.B., Torres, D.F.M.: Euler-Lagrange equations for composition functionals in calculus of variations on time scales. Discrete Contin. Dyn. Syst. **29**(2), 577–593 (2011)
46. Malinowska, A.B., Torres, D.F.M.: A general backwards calculus of variations via duality. Optim. Lett. **5**(4), 587–599 (2011)
47. Malinowska, A.B., Torres, D.F.M.: Introduction to the Fractional Calculus of Variations. Imperial College Press, London (2012)
48. Malinowska, A.B., Torres, D.F.M.: Quantum Variational Calculus. Springer Briefs in Electrical and Computer Engineering. Springer, Cham (2014)
49. Malinowska, A.B., Odzijewicz, T., Torres, D.F.M.: Advanced Methods in the Fractional Calculus of Variations. Springer Briefs in Applied Sciences and Technology. Springer, Cham (2015)
50. Martins, N., Torres, D.F.M.: Calculus of variations on time scales with nabla derivatives. Nonlinear Anal. **71**(12), e763–e773 (2009)

51. Martins, N., Torres, D.F.M.: Generalizing the variational theory on time scales to include the delta indefinite integral. Comput. Math. Appl. **61**(9), 2424–2435 (2011)
52. Martins, N., Torres, D.F.M.: Higher-order infinite horizon variational problems in discrete quantum calculus. Comput. Math. Appl. **64**(7), 2166–2175 (2012)
53. Martins, N., Torres, D.F.M.: Necessary optimality conditions for higher-order infinite horizon variational problems on time scales. J. Optim. Theory Appl. **155**(2), 453–476 (2012)
54. Mayer, A.: Die existenzbedingungen eines kinetischen potentiales. Math-Phys. Kl. **84**, 519–529 (1896)
55. Merrell, E., Ruger, R., Severs, J.: First order recurrence relations on isolated time scales. Panamer. Math. J. **14**(1), 83–104 (2004)
56. Odzijewicz, T., Torres, D.F.M.: The generalized fractional calculus of variations. Southeast Asian Bull. Math. **38**(1), 93–117 (2014)
57. Orlov, I.V.: Inverse extremal problem for variational functionals. Eurasian Math. J. **1**(4), 95–115 (2010)
58. Orlov, I.V.: Elimination of Jacobi equation in extremal variational problems. Methods Funct. Anal. Topology **17**(4), 341–349 (2011)
59. Saunders, D.J.: Thirty years of the inverse problem in the calculus of variations. Rep. Math. Phys. **66**(1), 43–53 (2010)
60. Segi Rahmat, M.R.: On some (q, h)-analogues of integral inequalities on discrete time scales. Comput. Math. Appl. **62**(4), 1790–1797 (2011)
61. Torres, D.F.M.: Proper extensions of Noether's symmetry theorem for nonsmooth extremals of the calculus of variations. Commun. Pure Appl. Anal. **3**(3), 491–500 (2004)
62. Torres, D.F.M.: The variational calculus on time scales. Int. J. Simul. Multidisci. Des. Optim. **4**(1), 11–25 (2010)
63. van Brunt, B.: The Calculus of Variations. Universitext, Springer, New York (2004)
64. Volterra, V.: Leçons sur les fonctions de lignes, Gauthier-Villars, Paris (1913)
65. Wyrwas, M.: Introduction to Control Systems on Time Scales. Lecture Notes, Institute of Cybernetics, Tallinn University of Technology (2007)

An Alternative Method for Snow Cover Mapping on Satellite Images by Modern Applied Mathematics

Semih Kuter, Zuhal Akyürek, Nazan Kuter and Gerhard-Wilhelm Weber

Abstract Continuous monitoring of snow cover is very crucial since the extend and amount of snow are key parameters for many processes closely related to ecology and climatology. Measuring the extend and amount of snow by in situ measurements are not always practical and possible due to operational and logistic reasons. Since 1960s, images taken by earth-observing satellites have been extensively used to monitor snow cover, and many parametric and nonparametric image classification methods have been proposed and applied for snow cover mapping. In this study, a novel application of nonparametric regression splines is introduced within the frame of modern applied mathematics and remote sensing. Implementation of *multivariate adaptive regression splines* (MARS) in image classification for snow mapping on *moderate resolution imaging spectroradiometer* (MODIS) images is demonstrated within a well-elaborated framework. The relation between the variations in MARS model building parameters and their effect on the predictive performance are represented in various perspectives. Performance of MARS in image classification is compared

S. Kuter (✉)
Faculty of Forestry, Department of Forest Engineering, Çankırı Karatekin University, 18200 Çankırı, Turkey
e-mail: semihkuter@yahoo.com

S. Kuter · G.-W. Weber
Institute of Applied Mathematics, Middle East Technical University,
06800 Ankara, Turkey
e-mail: gweber@metu.edu.tr

Z. Akyürek
Faculty of Engineering, Department of Civil Engineering, Middle East
Technical University, 06800 Ankara, Turkey
e-mail: zakyurek@metu.edu.tr

Z. Akyürek · G.-W. Weber
Graduate School of Natural and Applied Sciences, Department of Geodetic
and Geographic Information Technologies, Middle East Technical University,
06800 Ankara, Turkey

N. Kuter
Department of Landscape Architecture, Çankırı Karatekin University
Faculty of Forestry, 18200 Çankırı, Turkey

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_13

with the traditional *maximum-likelihood* (ML) method by using error matrices. Significant improvement in the classification accuracy of MARS models is observed as the number of basis functions and the degree of interaction increase. On three image sets out of four, the MARS approach gives better classification accuracies when compared to ML method.

Keywords Nonparametric regression splines · MARS · Snow cover mapping · Image classification · MODIS

1 Introduction

Snow is an important land cover on the earth's surface, and its distribution in space and time is a crucial parameter for a wide variety of reasons [1]. Snow cover not only plays a significant role on the hydrologic cycle and climatology of the earth, but also has an enormous influence on society and economical development [2].

Since snow has a high heat capacity, the soil surface is insulated from the atmosphere by the snow cover, and this slows down the warming process in spring. Hence, snow has a significant impact on micro and macro atmospheric circulation models as it affects energy absorption and thermal heating of the basin. In the process of heat and water vapor exchange between earth and atmosphere, one of the most important variables is snow cover [3]. Therefore monitoring of seasonal snow cover is a strict necessity in order to deepen our understanding for present and future climate, water cycle, and ecological changes [4].

It is almost impossible and impractical to map the snow cover by in situ measurements due to the extremely high cost and manpower required [1]. *Remote sensing* (RS) data available from various kinds of coarse and medium spatial resolution instruments is a powerful alternative, and has been employed to provide environmental data worldwide. Along with the parallel developments in the RS technologies, significant progress has been made in monitoring the snow cover since the mid-60s, when the first operational snow mapping was done by National Oceanic and Atmospheric Administration of U.S. [5].

As a passive RS instrument, *moderate resolution imaging spectro-radiometer* (MODIS) is probably one of the most frequently used instruments in snow cover mapping with its 36 spectral bands ranging in wavelength from 0.4 to 14.4 μm at varying spatial resolutions (bands 1–2: 250 m, bands 3–7: 500 m, and bands 8–36: 1,000 m) [6]. Since its launch in 1999, data collected by MODIS on the Terra satellite have been extensively used for mapping global snow cover through the snow mapping algorithm, where each MODIS 500-m pixel is classified as snow or non-snow [7]. The snow mapping algorithm is mainly based on the *normalized difference snow index* (NDSI), in which the MODIS bands 4 (centered at 0.555 μm) and 6 (centered at 1.640 μm) are used, along with a series of threshold tests and the MODIS cloud mask [8].

Apart from snow mapping algorithm developed for MODIS, many other techniques for snow cover mapping have been proposed and applied on data collected by various RS instruments with different technical and operational characteristics. These techniques can be grouped under two main categories: (*i*) parametric, and (*ii*) nonparametric ones. Our intend in this chapter is not enumerating these techniques in a detailed manner; however, one should note that each has its own drawbacks arising mainly from the inherent characteristic of multispectral image data.

In RS, individual pixels each carrying the radiometric information (i.e., reflectance or radiance values recorded by the sensor within different pre-defined spectral band intervals) are used to form a multispectral image [9]. The user should pay special attention when classifying multispectral images acquired by RS devices since they are very complex entities including not only spectral attributes but also spatial attributes. The phenomenon known as the *mixed pixel problem* [10] can often be encountered in real-life applications, especially, in complex classification problems, where classes are usually overlapping *spatially* and *spectrally*. The former means that a pixel can be contained in areas which are represented by more than one class, whereas the latter indicates that the radiometric characteristics of different classes may exhibit similar behaviors.

Parametric methods are mostly based on Bayesian approach, in which the training data are employed to estimate the parameters (i.e., mean and covariance matrices) of the probability density function of each class in order to generate the decision boundaries. These probability density functions are generally assumed to follow a normal multivariate distribution; however, this condition is hardly met in remotely sensed data [11]. Additionally, in case of complex landscapes with high-dimensional data, the number of samples that the user has to collect in order to train the classifier should be increased [11]. *Maximum likelihood* (ML) [12] is probably the most widely known parametric image classification method. Even though it is limited in solving the mixed pixel problem, it was one of the main classification methods used in RS until the mid-90s [9].

On the other hand, nonparametric methods are those that do not make any statistical a priori assumption about the underlying density function. Consequently, a nonparametric classifier does not need any statistical parameters, and the decision boundaries in a multidimensional feature space are obtained from training data of all classes [9]. As it has been shown by [13, 14], the performance of nonparametric methods, even with small training samples, are better than parametric ones. Some of the nonparametric approaches commonly used in image classification such as *artificial neural notworks* (ANN) [15] and *fuzzy classification* (FC) [16] methods suffer from certain problems related to their black box nature (i.e., no physical insight is available or used, but the chosen model is known to have good flexibility and has performed well in the past [17]). So, in such methods, it is hardly possible to understand the relation between the predictor variables and the classification results, which hinders the generalizability of the classification.

Nonparametric regression and classification techniques are mostly the key data mining tools in explaining real-life problems and natural phenomena where many effects often exhibit a nonlinear behavior. As an innovative nonparametric regression and classification tool, *multivariate adaptive regression splines* (MARS) [18] is a widely used algorithm in data mining and estimation theory in order to built flexible models for high-dimensional and nonlinear data. MARS is an evolved form of linear models that can automatically model nonlinear and interactive models, and it has great importance in both classification and regression.

In MARS model building, piecewise linear *basis functions* (BFs) are fitted in such a way that additive and interactive effects of the predictors are taken into account to determine the response variable. MARS uses two stages when building up a regression model, namely, the *forward* and the *backward step* algorithms. In the forward step, BFs are added up until the highest level of complexity is reached. Since the first step creates an over-fit model, preferred and eventual model is obtained by the elimination of BFs in the backward step. Selection of BFs is data-based and specific to the problem in MARS, which makes it an adaptive regression procedure suitable for solving high-dimensional problems [19].

MARS has many successful applications in various fields of science and engineering such as operational research, marketing and finance [20, 21]; ecology and forestry [22, 23]; simulation and computation [24]; geophysics [25]; engineering [26, 27]; medical sciences [28]; as well as in RS [29–31]. However, its implementation in RS for multispectral image classification is quite rare.

The study by Quirós et al. [32] is the only available one that used MARS as an alternative method in multispectral image classification. This study used an *advanced spaceborne thermal emission and reflection radiometer* (ASTER) image of a part of Spanish province of Badajoz. It has an area of 60 × 60 km, taken on 4 August 2000. The forest stand map in vector format that covers some of the study area was used for both training and testing. Total 17 land classes were a priori determined. Training polygons (they are often called *region of interest*, i.e., ROI), areas in which the training data were collected, were defined in the forest stand map by buffer analysis. Necessary training data were obtained by superimposing the ROIs for each class onto the satellite image, and then extracting the necessary pixel reflectance values for each spectral band.

The MARS classification on the image was carried out in binary fashion, i.e., each time, one of the classes was fixed and labeled as “1”, and the rest was considered belonging to the other class, so labeled as “0”. In this way, class probability maps were generated for each class, and then they were combined to obtain the final classification result. The result of the MARS classification was then compared with that of ML and parallelepiped classification methods in terms of *area under the curve* (AUC) statistics, and as the results revealed, 14 out of 17 classes, MARS gave better classification accuracies.

Even though this study revealed the potential of MARS algorithm for the classification of multispectral satellite images, it lacks certain important aspects, in our opinion:

- Total area of the ROIs (976 km^2), where the training samples were taken, was nearly 64% of the area used for accuracy assessment (1,523 km^2), which can easily be assumed as over safe,

- MARS algorithm allows the user to set certain model building parameters that directly affects the model's predictive performance such as the maximum number of BFs and the degree of interaction between predictor variables. However, this important issue and its impact on the final classification were not addressed,
- The binary classification approach used in the study is a bit over complex and time consuming in operational perspective, especially, when the number of classes is high,
- All the analysis covered a single geographic area, and were carried out on a single image taken by a medium spatial resolution sensor (i.e., ASTER). So, it is really difficult to reach a fair conclusion about the classification performance of MARS. We strongly believe that it should also be tested on images of different geographic regions taken by coarse resolution sensors such as MODIS, where the effect of mixed pixel can easily prevail due to the heterogeneous structure of landscapes over large areas.

As a result, our main goal in this chapter is to clarify the above mentioned shortcomings by introducing an image classification scheme for snow mapping via multiresponse MARS approach and evaluating its classification performance within a more comprehensible and well-established framework. For this purpose, four MODIS images, two of them taken over Alps and the other two over Turkey, are used as data set. Then, by systematically varying the model building parameters, different MARS classification models are obtained, and then applied on the images. Traditional ML classification approach is also used on the same data set. The performance of both approaches is compared by using error matrices and the effect of employing different model building parameters in the predictive performance of MARS is introduced based on the experimental findings. The remainder of this chapter is organized as follows. The next section gives a basic overview on MARS method, MODIS snow algorithm and MODIS daily snow product. The data set, the methodology and the results are introduced in Sect. 3. Finally, Sect. 4 wraps up the chapter by representing the conclusions, overall findings of the study and an outlook.

2 MARS Method, MODIS Snow Algorithm and MODIS Daily Snow Product

In this section, readers can find the basics of MARS and MODIS snow algorithms as well as a brief description of MODIS daily snow product.

2.1 MARS Method

The insight to the MARS approach in this subsection is given based on [18, 19, 31, 33].

In MARS, piecewise linear BFs are used in order to define relationships between a response variable and a set of predictors. The range of each predictor variable is cut into subsets of the full range by using knots which defines an inflection point along the range of a predictor. The slope of the linear segments between each consecutive pair of knots varies which ensures the fully-fitted function has no breaks or sudden steps.

The truncated piecewise linear BFs of MARS have the following form:

$$[\tilde{x}-\tau]_+ = \begin{cases} \tilde{x}-\tau, & \text{if}\, \tilde{x} > \tau, \\ 0, & \text{otherwise,} \end{cases}$$

$$[\tau-\tilde{x}]_+ = \begin{cases} \tau-\tilde{x}, & \text{if}\, \tilde{x} < \tau, \\ 0, & \text{otherwise.} \end{cases}$$

where $\tilde{x},\ \tau \in \mathbb{R}$. These two functions are also known as a *reflected pair* and the symbol '+' indicates that only the positive parts are used.

The following general model gives the relation between the predictor variables and their corresponding response:

$$Y = f(\tilde{\boldsymbol{X}}) + \varepsilon, \tag{1}$$

where response variable is indicated by Y, $\tilde{\boldsymbol{X}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p)^T$ is a vector of predictors and ε is an additive stochastic component having zero mean and finite variance. Then, MARS generates reflected pairs for each input $\tilde{X}_j$ $(j = 1, 2, \ldots, p)$ with p-dimensional knots $\boldsymbol{\tau}_i = (\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,p})^T$ at or just nearby each input data vectors $\tilde{\boldsymbol{X}}_i = (\tilde{x}_{i,1}, \tilde{x}_{i,2}, \ldots, \tilde{x}_{i,p})^T$ of that input $i = (1, 2, \ldots, N)$.

Then, the set of 1-dimensional BFs of MARS can be expressed as follows:

$$C := \big\{[x_j - \tau]_+, [\tau - x_j]_+ \mid \tau \in \{x_{1,j}, x_{2,j}, \ldots, x_{N,j}\},\ j \in \{1, 2, \ldots, p\}\big\}, \tag{2}$$

where N is the number of observations, p is the dimension of the input space. Here, $f(\tilde{\boldsymbol{X}})$ in Eq. 1 can be represented as a linear combination, which is successively constructed by the set C and with the intercept β_0, and it is in the following form:

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(\tilde{\boldsymbol{X}}^m) + \varepsilon, \tag{3}$$

where B_m is a BF or product of two or more BFs from the set C, and it is taken from a set of M linearly independent BFs. Here, $\tilde{\boldsymbol{X}}^m$ is a subvector of $\tilde{\boldsymbol{X}}$ contributing to the function B_m, and β_m denotes the unknown coefficient of the mth BF, or the constant 1 ($m = 0$). By multiplying an existing BF with another reflected pair including another variable, a new BF is generated that represents interaction between different variables, and both the existing BFs and the newly created BFs are included in the model. By this way, spline fitting in higher dimensions is achieved, leading to multivariate spline BFs with the following form:

$$B_m(\tilde{\boldsymbol{X}}^m) =: \prod_{j=1}^{K_m} [s_{\kappa_j^m} . (\tilde{x}_{\kappa_j^m} - \tau_{\kappa_j^m})]_+, \tag{4}$$

where the total number of truncated linear functions multiplied in the mth BF is denoted by K_m, $\tilde{x}_{\kappa_j^m}$ indicates the input variable corresponding to the jth truncated linear function in the mth BF, $\tau_{\kappa_j^m}$ is the knot location for $\tilde{x}_{\kappa_j^m}$, and finally $s_{\kappa_j^m} \in \{\pm 1\}$.

In order to estimate β_0, forward step algorithm of MARS starts with the constant function $B_0(\tilde{X}^0) = 1$. All functions from the set C are considered as candidate functions, and possible BFs have the following form:

- 1,
- $\tilde{x}_k$,
- $[\tilde{x}_k - \tau_i]_+$,
- $\tilde{x}_k \tilde{x}_l$,
- $[\tilde{x}_k - \tau_i]_+ \tilde{x}_l$,
- $[\tilde{x}_k - \tau_i]_+ [\tilde{x}_l - \tau_j]_+$.

MARS algorithm does not allow self-interaction between predictor variables (i.e., predictor variables cannot be the same for each BF). Therefore, x_k and x_l in the above BFs represent distinct predictor variables, together with their corresponding knot locations τ_i and τ_j, respectively. At each step, with one of the reflected pair in the set C, all products of a function $B_m(\tilde{\boldsymbol{X}}^m)$ in the model set are considered as a new function pair having the following form:

$$\beta_{M+1} B_k(\tilde{\boldsymbol{X}}^k) \cdot [\tilde{X}_j - \tau]_+ + \beta_{M+2} B_k(\tilde{\boldsymbol{X}}^k) \cdot [\tau - \tilde{X}_j]_+,$$

where β_{M+1} and β_{M+2} are coefficients estimated by least squares. For instance, the following BFs are potential candidates:

- 1,
- $\tilde{x}_k$,
- $[\tilde{x}_k - \tau_i]_+$, if $\tilde{x}_k$ is already in the model,
- $\tilde{x}_k \tilde{x}_l$, if $\tilde{x}_k$ and $\tilde{x}_l$ are already in the model,
- $[\tilde{x}_k - \tau_i]_+ \tilde{x}_l$, if $\tilde{x}_k \tilde{x}_l$ and $[\tilde{x}_k - \tau_i]_+$ are already in the model,
- $[\tilde{x}_k - \tau_i]_+ [\tilde{x}_l - \tau_j]_+$, if $[\tilde{x}_k - \tau_i]_+ \tilde{x}_l$ and $[\tilde{x}_l - \tau_j]_+ \tilde{x}_k$ are already in the model.

At each step, the algorithm chooses the knot and its corresponding pair of BFs that result in the largest decrease in residual error, and the products satisfying the above mentioned condition are successively added to the model until a user-defined value M_{max} is reached.

At the end, a *maximal* model that typically *overfits* the initial data is obtained. Then, the backward step is applied in order to prevent the model obtained in the forward step from over-fitting by decreasing the complexity of the model without degrading the fit to the data. It removes the BFs that give the smallest increase in the residual sum of squares at each step, which means that a predictor variable can be completely excluded from the model unless any of its BFs has a meaningful contribution to the predictive performance of the model, and this iterative procedure continues until an optimal number of effective terms are represented in the final model.

Among the sequence of models obtained from the above mentioned process, an estimated best model, $\hat{f}_\alpha$, with the optimum number of terms α that gives the best predictive fit is chosen through a *lack-of-fit* (LOF) criteria defined by *generalized cross-validation* (GCV), which is expressed as follows:

$$\mathrm{LOF}(\hat{f}_\alpha) = \mathrm{GCV}(\alpha) := \frac{\sum_{i=1}^{N}(Y_i - \hat{f}_\alpha(\tilde{X}_i))^2}{(1 - Q(\alpha)/N)^2}, \tag{5}$$

where N is the number of sample observations, $Q(\alpha) = u + dK$ with K representing the number of knots which are selected in forward pass and u is the number of linearly independent functions in the model, d denotes a cost for each BF optimization, and usually $d = 3$ ($d = 2$ is used when the model is additive). The numerator is the conventional residual sum of squares, which is penalized by the denominator that accounts for the increasing variance in case of increasing model complexity, i.e., while larger $Q(\alpha)$ creates a smaller model with less number of BFs, smaller $Q(\alpha)$ generates a larger model with more BFs. Using the *lack-of-fit* criteria, the best model is chosen according to backward pass that minimizes GCV.

2.2 MODIS Snow Algorithm and MODIS Daily Snow Product

In this subsection, a brief overview on MODIS snow algorithm and MODIS daily snow product (i.e., MOD/MYD10A1) is introduced based on [1, 3, 6, 19].

Main logic behind snow detection is the fact that the reflectance of snow is high in the visible and low in the near infrared region. MODIS uses a fully automated snow mapping algorithm, in which bands 4 (0.545–0.565 μm) and 6 (1.628–1.652 μm) are used to calculate the NDSI value as given in the following equation:

$$\text{NDSI} = \frac{\text{band}_4 - \text{band}_6}{\text{band}_4 + \text{band}_6}. \tag{6}$$

If the NDSI of a pixel in a non-densely forested area is ≥ 0.4 and its reflectance in band 2 (0.841–0.876 µm) is >11%, it is labeled as snow. However, very low reflectance values make the denominator of the NDSI considerably small, and pixels with very dark targets, like black spruce forests, can be erroneously classified as snow. Therefore, when the reflectance in band 4 (0.545–0.565 µm) is <10%, then the pixel is not labeled as snow even if the other conditions are met.

When a forest stand is covered by snow, its spectral response changes in such a way that reflectance in the visible generally increases with respect to near infra-red reflectance, which results in a decrease in the *normalized difference vegetation index* (NDVI). In order to improve the snow mapping in dense forests, NDSI is used together with NDVI. MODIS bands 1 (0.620–0.670 µm) and 2 (0.841–0.876 µm) are employed in the calculation of NDVI, and if its value ≈ 0.1, the pixel can be labeled as snow even if the value of NDSI is < 0.4.

Additionally, MODIS infra-red bands 31 (10.780–11.280 µm) and 32 (11.770–12.270 µm) are employed via a split-window technique to estimate the ground temperature in order to eliminate the spurious snow and to increase the accuracy of snow mapping. By using this thermal mask, pixels with temperature >283 K are not labeled as snow.

There are seven MODIS snow products produced at Level 2 or Level 3, and they are at different temporal and spatial resolutions (cf. Table 1). The file format for snow products is HDF-EOS (i.e., *hierarchical data format—earth observing system*). Each daily snow product, MOD/MYD10A1, is a $10^0 \times 10^0$ tile (1200 × 1200 km) with a sinusoidal projection. Four scientific data sets are available in a MOD/MYD10A1 product: *snow cover map*, *fractional snow cover*, *snow albedo*, and finally, *quality assurance* (QA) data. In order to select an observation for the day, the scoring algorithm given in the following equation is employed:

$$\text{score} = 0.5SE + 0.3ND + 0.2OC, \tag{7}$$

where SE indicates the solar elevation, ND is the distance from nadir, and OC is the observation coverage. In order to generate the snow cover map, observation closest to local noon time (i.e., highest solar elevation angle) nearest to nadir with the greatest coverage is selected by the scoring algorithm, and then classified as snow, snow-covered water bodies, land, water, cloud or as other condition.

Observations from the fractional snow cover of L2G product are used to determine the daily fractional snow cover, again by using the same scoring algorithm in Eq. 7. Fractional snow is given within 0–100% range by the fractional snow cover map. This map includes inland water bodies, and non-snow pixels are classified as water, cloud or other condition.

By using the MODIS surface reflectance product, the snow albedo is determined for the visible and near infra-red bands. The resultant map shows the snow albedo

Table 1 Seven MODIS snow data products

Name of the product	Level	Data dimension	Spatial resolution	Temporal resolution
MOD/MYD10L2	L2	1354 × 2000 km	500 m	Swath
MOD/MYD10L2G	L2G	1200 × 1200 km	500 m	Day of multiple coincident swaths
MOD/MYD10A1	L3	1200 × 1200 km	500 m	Day
MOD/MYD10A2	L3	1200 × 1200 km	500 m	8 days
MOD/MYD10C1	L3	$360^0 \times 180^0$ (global)	$0.5^0 \times 0.5^0$	Day
MOD/MYD10C2	L3	$360^0 \times 180^0$ (global)	$0.5^0 \times 0.5^0$	8 days
MOD/MYD10C	L3	$360^0 \times 180^0$ (global)	$0.5^0 \times 0.5^0$	Month

Table 2 MOD/MYD10A1 daily snow tile attributes

Coordinate system	Cartesian
Valid range	0 – 254
Fill value	255 (used to fill gaps in the swath)
Key to data values	0: missing data
	1: no decision
	11: night
	25: no snow
	37: lake
	39: ocean
	50: cloud
	100: lake ice
	200: snow
	254: detector saturated
	255: fill

within the range of 0–100, and non-snow features are labeled with different values. A summary of attributes in a MOD/MYD10A1 daily snow tile is given in Table 2.

3 Data Set, Methodology and Results

In this section, the readers can find the details of the data set and employed methodology as well as the obtained results.

Table 3 MODIS data set used in the analysis (A: Alps, T: Turkey)

Data set	Date	Image product type	Image size (row × column)
A1	10.03.2002	MOD02HKM	3423 × 5713
		MOD09GA	2176 × 3384
		MOD10A1	2176 × 3384
A2	13.01.2006	MOD02HKM	3409 × 5822
		MOD09GA	2176 × 3384
		MOD10A1	2176 × 3384
T1	18.12.2006	MOD02HKM	3550 × 5703
		MOD09GA	4790 × 9372
		MOD10A1	4790 × 9372
T2	22.03.2009	MOD02HKM	3500 × 5699
		MOD09GA	4790 × 9372
		MOD10A1	4790 × 9372

3.1 Image Set and Model Training

The data set used in the study consists of four MODIS Terra images taken over two different geographic regions: Alps and Turkey. This set comprises three different products for each date, calibrated earth view image (MOD02HKM), surface reflectance image (MOD09GA), and MOD10A1 daily snow cover image. Readers can find comprehensive information on MOD2HKM, MOD09GA and MOD10A1 products in [6]. All images have 500 m spatial resolution. The details of the images are given in Table 3.

In model training phase, reflectance values from the solar reflective bands (bands 1–7) in MOD02HKM images are used as predictor variables. By using the 1st, 3rd and 4th bands from MOD09GA images, an RGB color composite image of each area is obtained (R = 1st, G = 4th, and B = 3rd band). On each image, a test area is defined (cf. Fig. 1) in ArcMap software [34].

The first step is to decide the number of classes in the images. MODIS is a coarse resolution instrument with *large-field-of-view* sensors, and its tiles cover large areas. Additionally, the test regions have surface areas changing roughly from 62,500 to 67,500 km^2, and it is hardly possible to find a reference data for accuracy assessment on such large areas. Therefore, the basic land cover types already available in MOD10A1 are chosen, namely, *snow*, *water*, *cloud*, and *land* (i.e., no snow) so that MOD10A1 can also be used in evaluating the performances of the classifiers.

In MARS model training, the *earth* module [35] under *R* statistical software [36] is used. The *earth* requires two matrices to build MARS models, the matrix of predictor variables (i.e., reflectance values of bands 1–7), and the matrix of responses (i.e., the pixel's corresponding class). Since we are dealing with a multiresponse MARS classification, instead of a binary classification as applied in [32], a special design

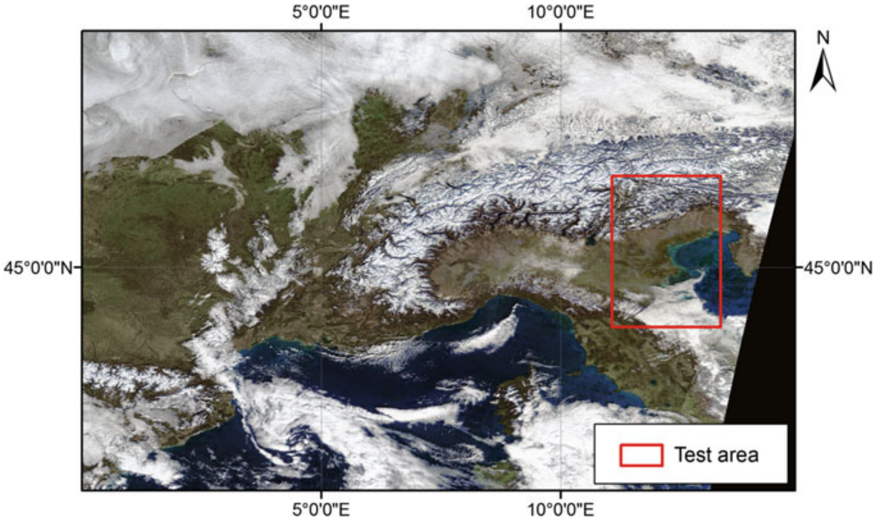


Fig. 1 MOD09GA RGB color composite image of A2

for the response matrix is necessary. Our response matrix have four columns, each of which represents one of the classes (i.e., snow, water, cloud, and land).

At each row, the corresponding class is assigned "1" under the associated column and the other cells are labeled as "0". By this way, *earth* produces four simultaneous models, each have the same set of BFs, but different coefficients. For each class, ROIs for training samples are marked on the associated RGB image in such a way that they never fall inside or overlap with the test area on that image (i.e., ROIs are excluded from the test areas), and they are saved as shape files in ArcMap (cf. Fig. 2).

In order to prepare the necessary response-predictor matrix pairs for each data set, reflectance values from bands 1–7, together with the associated class labels, are extracted onto the pixels delineated by ROIs by using a code written in MATLAB [37]. The results are saved in text files. The average percentage of training data to the associated test area is nearly 25%. The details on the test areas and the training data can be found in Table 4.

The predictor-response matrix pair of each data set is introduced into *R*. Then the MARS classification models are generated for different settings of model building parameters (i.e., the maximum numbers of BFs and the degree of interaction between predictor variables). For all settings, threshold for stopping criteria is 10^{-6}. These settings are given in Table 5.

Multiresponse MARS implementation in *earth* module under *R* allows to fit multiple response variables in exactly the same way as a single-response MARS model. However, in multiresponse model fitting, residual squared errors are averaged across all response variables (i.e., classes). The obtained multiresponse model has a common set of BFs for all classes, yet uses a different set of coefficients for each class, which means the shapes of the fitted functions can differ between classes.

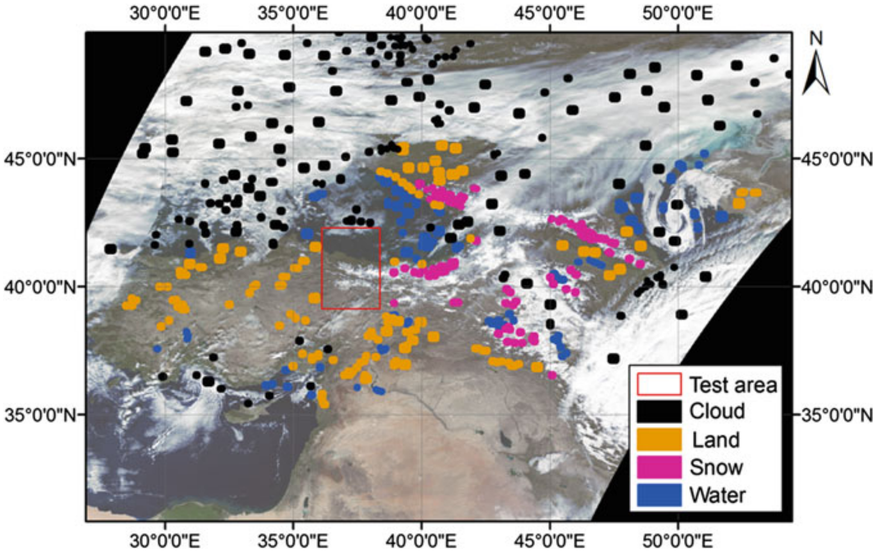


Fig. 2 Training samples for T1 image

Table 4 Training and test data details

Data set	Test image size(row × column)	Training data size (pixels)					
		Snow	Water	Cloud	Land	Total	% to test area
A1	689 × 494 = 340366 pixels	7796	17700	12154	40216	77866	22.9
A2	688 × 494 = 339872 pixels	6214	15800	26678	41300	89992	26.5
T1	758 × 544 = 412352 pixels	8065	16261	47555	32452	104333	25.3
T2	758 × 544 = 412352 pixels	9821	18031	36286	48732	112870	27.4

3.2 Testing of MARS Models

For each data set, the reference image of the related test area is generated from the associated MOD10A1 daily snow product. In MOD10A1 daily snow images, water mask is superimposed on all other classes, including cloud. Consequently, the cloud cover over water bodies, if exists any, is suppressed by the water mask. To deal with the issue, a cloud mask is generated from the QA data available in MOD09GA image by using a MATLAB code. The cloud mask generated for A2 is given in Fig. 3. The

Table 5 Settings applied in MARS model building

Setting	Degree of interaction	Max. number of BFs
S1	1	20
S2		40
S3		60
S4		80
S5		100
S6	2	20
S7		40
S8		60
S9		80
S10		100
S11	3	20
S12		40
S13		60
S14		80
S15		100

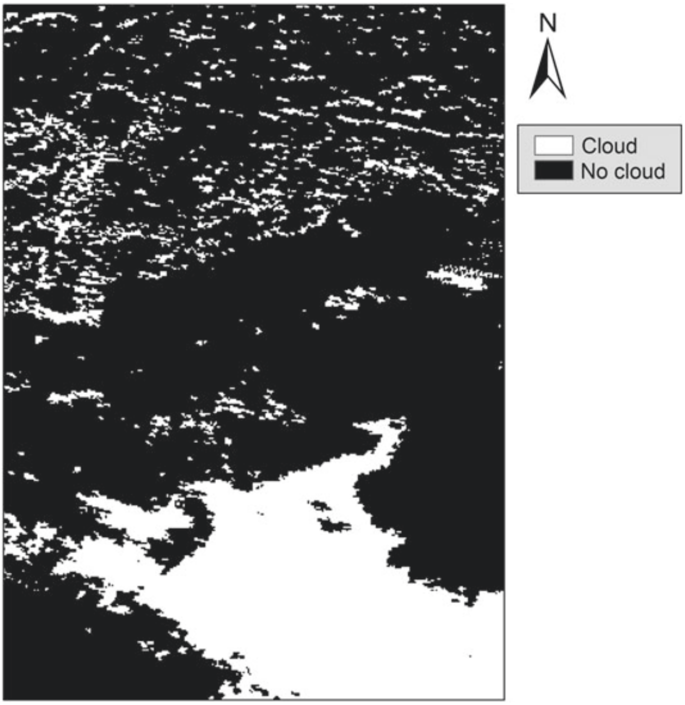


Fig. 3 Cloud mask generated from MOD09GA QA data for A1 test image

cloud mask for each image is resampled to 500 m spatial resolution, and then applied onto the corresponding test area in order to obtain the final reference image to be used in accuracy assessment.

In order to evaluate the performances of the MARS and ML classification approaches, the classified images are compared with the corresponding reference images by using error matrices.

3.2.1 Error Matrix

The classification accuracy of remotely sensed data is often expressed by an error matrix [38]. It is composed of square arrays of numbers arranged in the form of rows and columns. The cells in the matrix represent the number of pixels assigned to a particular class with respect to the actual class as verified in the reference data [39].

When the total number of correctly classified pixels in a class is divided by the total number of pixels of that class as derived from the reference data (i.e., the row total), *omission error* (i.e., producer's accuracy) is obtained. This measure gives the probability of a reference pixel being correctly classified.

The *commission error* (i.e., user's accuracy) is calculated by taking the ratio of the total number of correct pixels in a class to the total number of pixels that are classified in that class (i.e., the column total). Commission error refers to the probability that a pixel labeled as a certain class in the map is really this class.

Overall accuracy, the simplest descriptive statistic, is calculated by dividing the total correctly classified pixels (i.e., main diagonal) by the total number of pixels. An example of an error matrix is given in Table 6.

3.3 *Results and Discussion*

The highest overall classification accuracy obtained by MARS for each data set with the related model setting is given in Table 7 together with the corresponding ML classification result.

Table 6 A sample error matrix

		Classified data				
		Class 1	Class 2	Class 3	Class 4	Row total
Reference data	Class 1	260	16	88	96	460
	Class 2	24	324	20	32	400
	Class 3	0	44	340	76	460
	Class 4	16	28	12	360	416
	Column total	300	412	460	564	**868**

Table 7 The best overall accuracies obtained by MARS with corresponding MARS model building settings against ML method

Data set	MARS model setting	Overall accuracy by MARS (%)	Overall accuracy by ML (%)
A1	S15	94.387	92.918
A2	S13	82.753	82.711
T1	S15	82.704	82.706
T2	S12	86.979	84.624

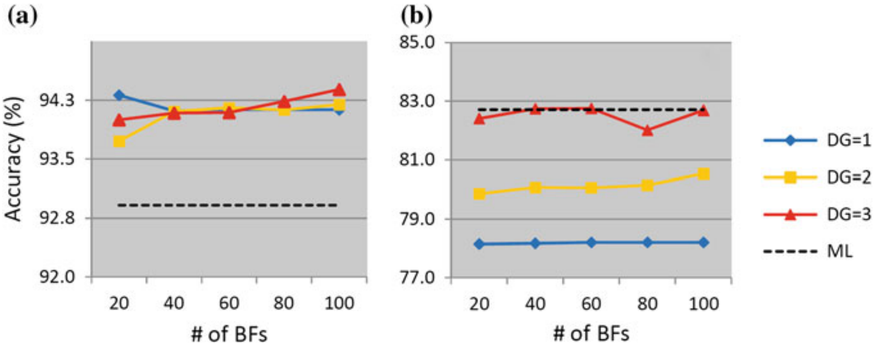


Fig. 4 Overall accuracies for **a** A1, and **b** A2 data sets

At first glance, MARS gives better overall accuracy for A1, A2 and T2 data sets. Even though the performance of ML method on T1 seems better than MARS, their overall accuracies are very close to each other.

In Figs. 4 and 5, the overall accuracies of MARS models with different degree of interactions (DG) are plotted against the number of BFs for each data set. When these figures are analyzed, it can be easily seen that when the degree of interaction is set to one (i.e., additive modeling) and the number of BFs is increased, no remarkable change in the overall accuracy is achieved for all data sets. In Figs. 4 and 5, black dashed line indicates the overall accuracy level obtained by ML classification for that particular data set. It should also be noted that the vertical axes of these graphs are not in common scale for better illustrative purpose.

For the second and the third degree of interactions in A1 and A2 data sets, it can also be concluded that when the degree of interaction is fixed, increasing the number of BFs may not always contribute significantly to the model's predictive performance. On the other hand, increase in both the number of BFs and the degree of interaction results in better classification performance as observed for T1 and T2 data sets.

In order to have more detailed understanding on the relation between the model building parameters and the predictive performance of MARS classification, it is also necessary to analyze the number of selected terms (ST) after the backward pass,

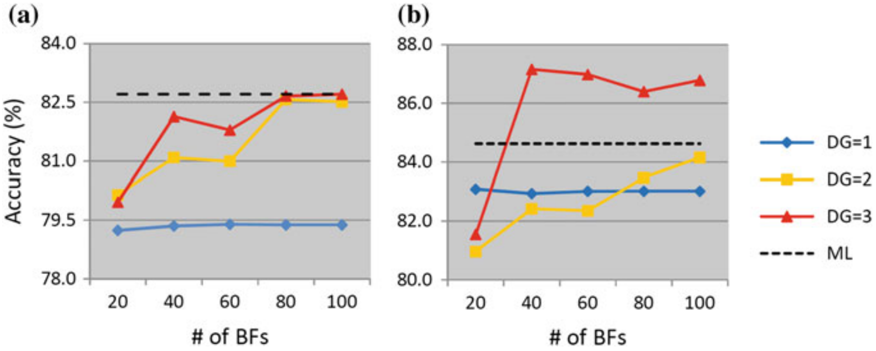


Fig. 5 Overall accuracies for **a** T1, and **b** T2 data sets

Table 8 Training performance of MARS models for A1 and A2 data sets

Setting	A1			A2		
	GCV	R^2	ST	GCV	R^2	ST
S1	0.04576	0.92940	17	0.03658	0.94510	15
S2	0.04371	0.93250	27	0.03571	0.94640	26
S3	0.04358	0.93280	37	0.03565	0.94660	33
S4	0.04356	0.93280	45	0.03564	0.94660	35
S5	0.04356	0.93280	45	0.03564	0.94660	35
S6	0.02945	0.95450	19	0.01614	0.97580	17
S7	0.02176	0.96650	34	0.01239	0.98140	29
S8	0.01944	0.97010	49	0.01085	0.98380	39
S9	0.01855	0.97150	62	0.01026	0.98460	53
S10	0.01784	0.97260	74	0.00987	0.98520	63
S11	0.03076	0.95250	17	0.01382	0.97930	16
S12	0.02089	0.96780	32	0.00972	0.98540	28
S13	0.01743	0.97310	43	0.00827	0.98760	39
S14	0.01567	0.97590	55	0.00733	0.98900	49
S15	0.01437	0.97790	65	0.00659	0.99010	59

R^2 and GCV values obtained during the model training phase, which are given in Tables 8 and 9.

It can easily be inferred from Tables 8 and 9 that GCV and R^2 values for S1, S2, S3, S4 and S5 (DG = 1) remain stable, supporting our conclusion. At these settings, although the number of terms included in the final model after the backward pass increase, no remarkable change in the GCV and R^2 values is seen. When the number of BFs is fixed and the degree of interaction is increased, as for S5, S10 and S15 in A2, significant change is achieved in GCV and R^2 values. This is a typical, and expected, characteristic of MARS method because higher settings for the number of

Table 9 Training performance of MARS models for T1 and T2 data sets

Setting	T1			T2		
	GCV	R^2	ST	GCV	R^2	ST
S1	0.05561	0.91650	16	0.04207	0.93790	16
S2	0.05409	0.91880	26	0.04075	0.93990	26
S3	0.05374	0.91930	36	0.04063	0.94010	37
S4	0.05364	0.91950	45	0.04061	0.94010	45
S5	0.05364	0.91950	47	0.04061	0.94010	45
S6	0.03279	0.95070	17	0.02651	0.96090	19
S7	0.02279	0.96580	33	0.02180	0.96780	36
S8	0.01923	0.97110	44	0.01960	0.97110	50
S9	0.01727	0.97410	55	0.01876	0.97240	65
S10	0.01663	0.97510	68	0.01847	0.97280	75
S11	0.03218	0.95170	16	0.02563	0.96220	19
S12	0.02147	0.96780	30	0.01852	0.97270	33
S13	0.01736	0.97390	44	0.01541	0.97730	45
S14	0.01450	0.97830	54	0.01381	0.97970	56
S15	0.01292	0.98060	64	0.01308	0.98070	66

BFs and the degree of interaction allow MARS to enlarge its search space in order to add more terms into the model, which increase its predictive ability.

It is also of value to mention that the reference data used in the accuracy assessment is MODIS daily snow product (i.e., MOD10A1). This product is composed of different layers of data obtained by various algorithms such as snow algorithm, land/water mask and cloud mask algorithms. Therefore, errors inherent with each layer also pass to MOD10A1 product, and can eventually have a negative contribution to our accuracy assessment. On the other hand, it is the best available product when the type of the classification scheme and the sizes of the test areas are considered.

Further comments on the results would be made on the classified images by visual analysis. The classified images of A1, A2, T1 and T2 data sets are given in Figs. 6, 7, 8 and 9, respectively.

In Fig. 6, some overlapping of cloud with land and water is apparent for ML method. In Fig. 7c, overlapping of water with snow for ML method is clearly seen. For MARS method (Fig. 7d), relatively slighter overlapping is observed between cloud, water and land.

For ML method in Fig. 8, relatively high overlapping between land and cloud exists, as compared to MARS method.

In Fig. 9, overall classification performance of MARS seems better than ML method. Small water body in the western part of Fig. 9b is not recognized by ML method, but successfully classified as water by MARS.

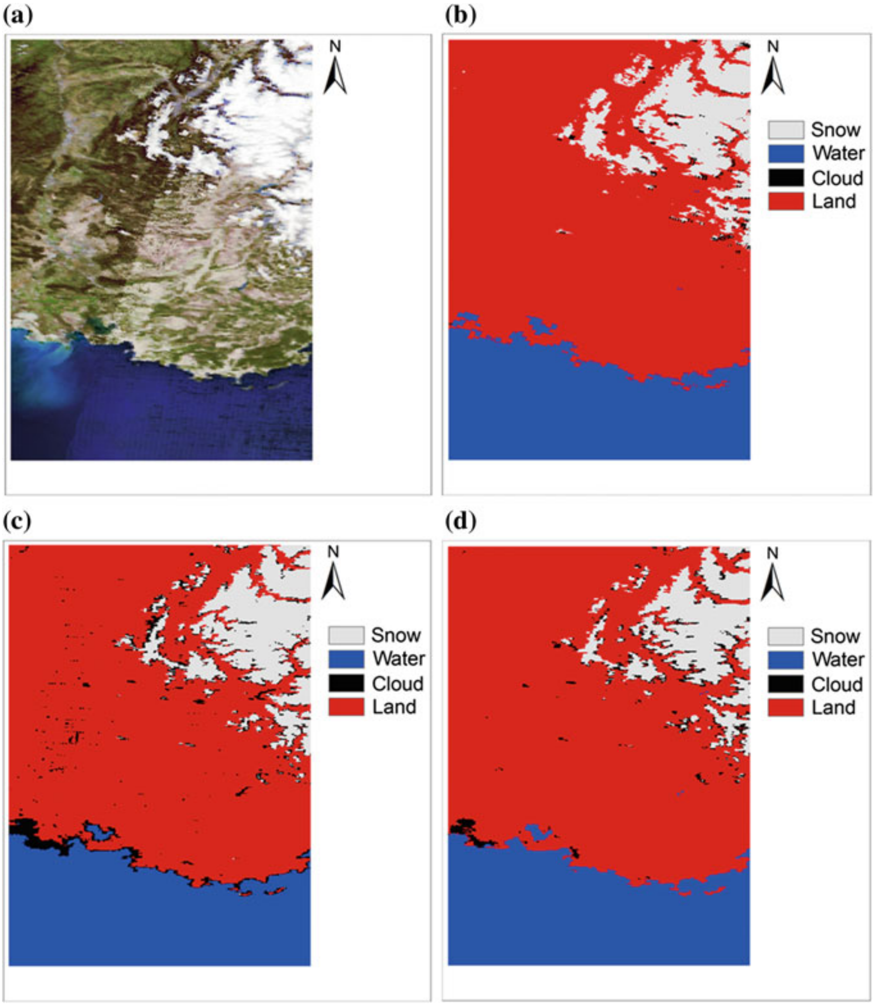


Fig. 6 Classified images of A1 data set: **a** MOD09GA RGB color composite image, **b** MOD10A1 reference image, **c** ML classification, and **d** MARS classification

According to our visual inspections on the classified images with respect to RGB and MOD10A1 images, MARS has relatively better performance than ML classifier for all data sets.

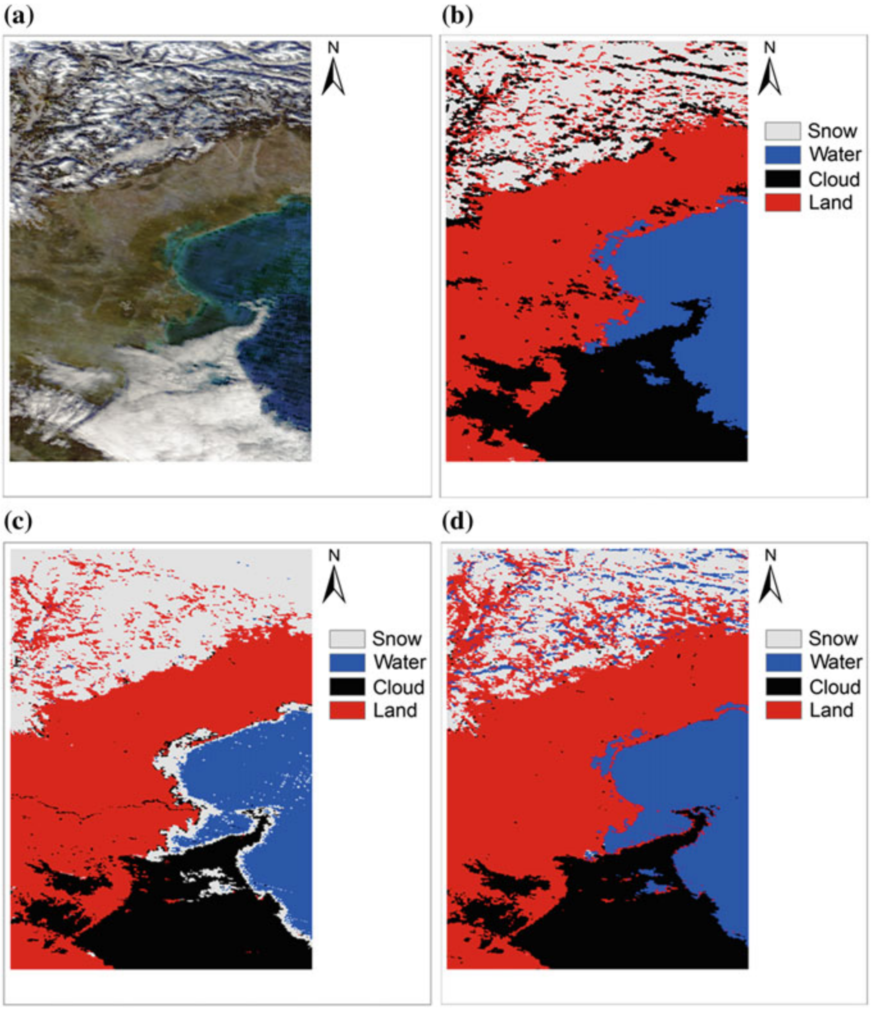


Fig. 7 Classified images of A2 data set: **a** MOD09GA RGB color composite image, **b** MOD10A1 reference image, **c** ML classification, and **d** MARS classification

4 Conclusions and Outlook

In this chapter, we have approached an active research topic in RS within a different and progressive perspective. We have dealt with image classification within the frame of nonparametric regression splines. This study demonstrates the applicability of multiresponse MARS approach for snow mapping on MODIS images by comparing the results with traditional ML classification method.

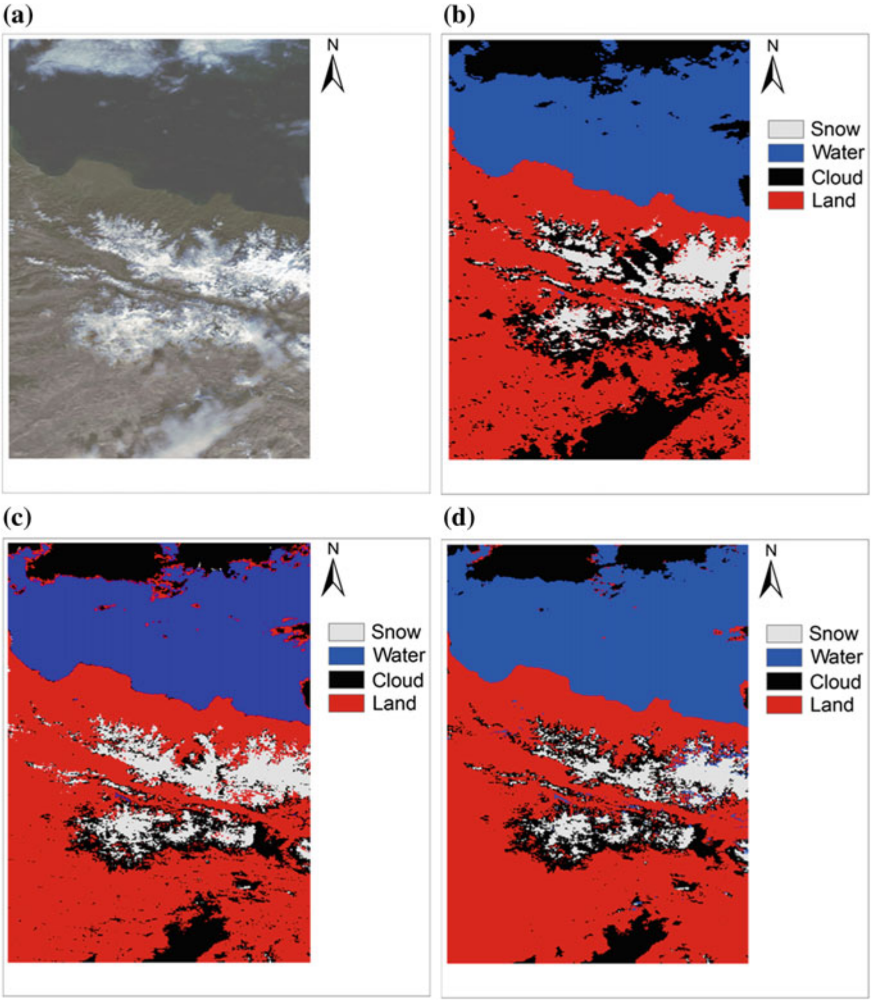


Fig. 8 Classified images of T1 data set: **a** MOD09GA RGB color composite image, **b** MOD10A1 reference image, **c** ML classification, and **d** MARS classification

According to the results, larger number of BFs and higher degree of interaction should be preferred for multispectral image classification by multiresponse MARS model. However, it should also be noted that there is no unique setting for the best model since the performance of the MARS is highly dependent on the training data. Therefore, the user should certainly "play" with the primary MARS model building parameters, i.e., the maximum number of BFs and the degree of interaction, in order to observe their effects on the model's behavior during the model training phase.

One drawback of building multiresponse MARS model by using *earth* module in *R* is that it does not allow users to adjust classification thresholds (i.e., cut-off values)

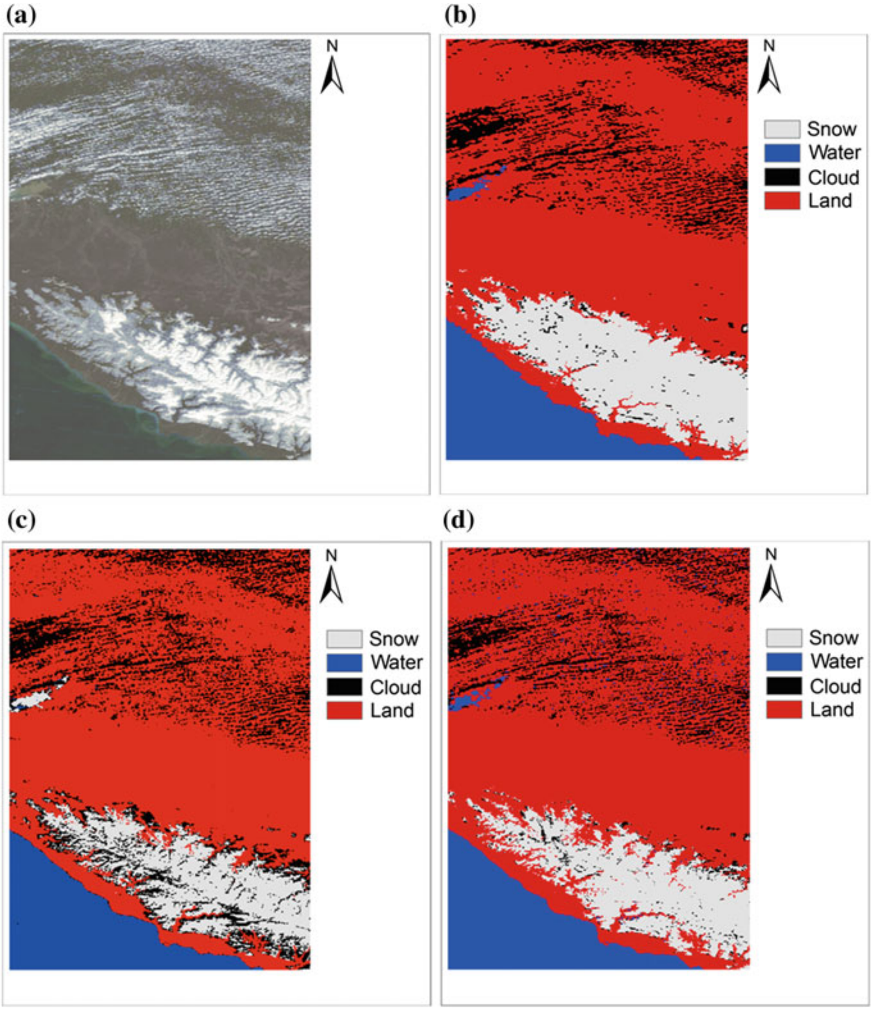


Fig. 9 Classified images of T2 data set: **a** MOD09GA RGB color composite image, **b** MOD10A1 reference image, **c** ML classification, and **d** MARS classification

for individual classes in a binary fashion. On the other hand, binary classification approach may not seem operationally practical if the number of classes is large.

Although other alternative nonparametric image classification techniques such as *support vector machines*, ANN and FC exist, some of these approaches are known to have problems in explaining the connection between the predictor variables and the results of classification. But MARS has a clear advantage over black-box approaches due to its well-elaborated statistical basis.

In this study, MARS approach has proven its competitiveness and applicability in image classification as compared to traditional Bayesian-based ML approach.

Within the light of the obtained results, it can also be concluded that other types of high-dimensional and complex RS data sets can be wisely handled with the inherent smoothing characteristic of splines.

Thus, we hope that the demonstrated MARS classification scheme and the represented results would serve as a guide and be helpful for those who will consider MARS for their image classification and other RS-related applications

In future, we plan to extend our study by also employing the optimization supported varieties of MARS, namely CMARS (i.e., Conic MARS) [40] and its robust counterpart RCMARS (i.e., Robust CMARS) [41], and to compare the results by our statistical performance criteria.

CMARS proposes an alternative formulation of the backward step of MARS by using the modern methods of *continuous optimization*. Instead of applying backward step algorithm, a *penalized residual sum of squares* is introduced to MARS as a *Tikhonov regularization* problem, and this two-objective optimization problem is treated using the continuous optimization technique called *conic quadratic programming*. CMARS provides a special advantage, namely its generalization in the form of RCMARS through the involvement of *robust optimization*. By this, the *regularization* in CMARS becomes rigorously extended towards a *robustification* which especially addresses uncertainty in the input variables, too, and goes beyond noise in the response variable by uncertainty.

RS technologies, together with the parallel developments in *geographical information systems* (GIS), have provided researchers and scientists with unique capabilities for editing, managing, analyzing and automating different kinds of spatial data as well as visualizing the spatial context that provides valuable information. Our ultimate aim is to integrate the dynamical progress of scientific advances in modern continuous optimization within the spatial technologies so that we can enhance our understanding of the value of spatial data and its inherent structure to reach better modeling capabilities in GIS and RS. The following two items can be proposed as the potential future extensions of our current study:

- Analyzing sequences of satellite images for observing landscape change and urban growth [42, 43] as well as urban green space networks for the identification of corridors that helps conserve biodiversity in urban areas threatened by urban sprawl [44] will certainly have great potential, which inherently implies analysis of dynamic systems and offers high applicability of our state-of-the-art nonparametric regression and classification tools, i.e., CMARS and RCMARS.
- Another interesting direction to pursue would be to employ our varieties CMARS and RCMARS of MARS within RS and GIS for the detection and image reconstruction of historical and archaeological sites [45, 46] as well as ancient road networks [47] in the Middle East and at further ancient regions of the world. By this, we aim at a new and helpful technology in landscape and settlement archaeology which could be competitive to given tools and allow conclusions about cultures and life in the past, about catastrophes which happened, improved early warning systems and a better preparation of humankind for the future.

References

1. Salomonson, V.V., Appel, I.: Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote Sens. Environ. **89**, 351–360 (2004)
2. Liang, T.G., Huang, X.D., Wu, C.X., Liu, X.Y., Li, W.L., Guo, Z.G., Ren, J.Z.: An application of MODIS data to snow cover monitoring in a pastoral area: a case study in Northern Xinjiang, China. Remote Sens. Environ. **112**, 1514–1526 (2008)
3. Tekeli, A.E., Akyürek, Z., Şorman, A.A., Şensoy, A., Şorman, Ü.: Using MODIS snow cover maps in modeling snowmelt runoff process in the eastern part of Turkey. Remote Sens. Environ. **97**, 216–230 (2005)
4. Czyzowska-Wisniewski, E.H., van Leeuwen, W.J.D., Hirschboeck, K.K., Marsh, S.E., Wisniewski, W.T.: Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network. Remote Sens. Environ. **156**, 403–417 (2015)
5. Gafurov, A., Brdossy, A.: Cloud removal methodology from MODIS snow cover product. Hydrol. Earth Sys. Sci. **13**, 1361–1373 (2009)
6. Qu, J.J., Gao, W., Kafatos, M., Murphy, R.E., Salomonson, V.V.: Earth Science Satellite Remote Sensing. Volume 1: Science and Instruments. Springer, Beijing (2006)
7. Salomonson, V.V., Appel, I.: Development of the aqua MODIS NDSI fractional snow cover algorithm and validation results. IEEE Trans. Geosci. Remote Sens. **44**, 1747–1756 (2006)
8. Hall, D.K., Riggs, G.A., Salomonson, V.V., DiGirolamo, N.E., Bayr, K.J.: MODIS snow-cover products. Remote Sens. Environ. **83**, 181–194 (2002)
9. Tso, B., Mather, P.M.: Classification Methods for Remotely Sensed Data, 2nd edn. CRC Press, Boca Raton (2009)
10. Ines, A.V., Honda, K.: On quantifying agricultural and water management practices from low spatial resolution RS data using genetic algorithms: a numerical study for mixed-pixel environment. Adv. Water Resour. **28**, 856–870 (2005)
11. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. **28**, 823–870 (2007)
12. Schowengerdt, R.A.: Remote Sensing: Models and Methods for Image Processing, 3rd edn. Academic Press, New York (2006)
13. Cortijo, F., De La Blanca, N.P.: The performance of regularized discriminant analysis versus non-parametric classifiers applied to high-dimensional image classification. Int. J. Remote Sens. **20**, 3345–3365 (1999)
14. Raudys, S.: On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 667–671 (1997)
15. Civco, D.L.: Artificial neural networks for land-cover classification and mapping. Int. J. Geogr. Inf. Sys. **7**, 173–186 (1993)
16. Foody, G.M.: Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. Int. J. Remote Sens. **17**, 1317–1340 (1996)
17. Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson, H., Juditsky, A.: Nonlinear black-box modeling in system identification: a unified overview. Automatica **31**, 1691–1724 (1995)
18. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**, 1–67 (1991)
19. Kuter, S.: Atmospheric correction and image classification on MODIS images by nonparametric regression splines (Ph.D. thesis), The Graduate School of Natural and Applied Sciences, Department of Geodetic and Geographic Information Technologies. Middle East Technical University, Ankara, Turkey (2014)
20. Özmen, A., Kropat, E., Weber, G.-W.: Spline regression models for complex multi-modal regulatory networks. Optim. Methods Softw. **29**, 515–534 (2014)
21. Alp, Ö.S., Büyükbebeci, E., Çekiç, Aİ., Özkurt, F.Y., Taylan, P., Weber, G.-W.: CMARS and GAM & CQP - modern optimization methods applied to international credit default prediction. J. Comput. Appl. Math. **235**, 4639–4651 (2011)
22. Özmen, A., Batmaz, İ., Weber, G.-W.: Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS. Environ. Model. Assess. **19**, 425–435 (2014)

23. Henne, P.D., Hu, F.S., Cleland, D.T.: Lake-effect snow as the dominant control of mesic-forest distribution in Michigan, USA. J. Ecol. **95**, 517–529 (2007)
24. Zhou, Y., Leung, H.: Predicting object-oriented software maintainability using multivariate adaptive regression splines. J. Sys. Softw. **80**, 1349–1361 (2007)
25. Krzyścin, J.W., Eerme, K., Janouch, M.: Long-term variations of the UV-B radiation over Central Europe as derived from the reconstructed UV time series. Ann. Geophys. **22**, 1473–1485 (2004)
26. Mukhopadhyay, A., Iqbal, A.: Prediction of mechanical property of steel strips using multivariate adaptive regression splines. J. Appl. Stat. **36**, 1–9 (2009)
27. Samui, P., Das, S., Kim, D.: Uplift capacity of suction caisson in clay using multivariate adaptive regression spline. Ocean Eng. **38**, 2123–2127 (2011)
28. Chou, S.M., Lee, T.S., Shao, Y.E., Chen, I.F.: Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Sys. Appl. **27**, 133–142 (2004)
29. Durmaz, M., Karslıoğlu, M.O., Nohutcu, M.: Regional VTEC modeling with multivariate adaptive regression splines. Adv. Space Res. **46**, 180–189 (2010)
30. Kuter, S., Weber, G.W., Özmen, A., Akyürek, Z.: Modern applied mathematics for alternative modeling of the atmospheric effects on satellite images. In: Pinto, A.A., Zilberman, D. (eds.) Modeling, Dynamics, Optimization and Bioeconomics I, pp. 469–485. Springer, Berlin (2014)
31. Kuter, S., Weber, G.-W., Akyürek, Z., Özmen, A.: Inversion of top of atmospheric reflectance values by conic multivariate adaptive regression splines. Inverse Probl. Sci. Eng. **23**, 651–669 (2015)
32. Quirós, E., Felicísimo, Á.M., Cuartero, A.: Testing multivariate adaptive regression splines (MARS) as a method of land cover classification of TERRA-ASTER satellite images. Sensors **9**, 9011–9028 (2009)
33. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Berlin (2009)
34. ArcMap^{TM}, ESRI ArcMap Version 9.3.1, 1999–2009 ESRI Inc
35. Milborrow, S.: Earth: Multivariate adaptive regression spline models - derived from mda:mars by Trevor Hastie and Rob Tibshirani, R package version 3.2-2. http://CRAN.R-project.org/package=earth 2012
36. R Development Core Team.: R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/. 2012
37. MATLAB®, R2012b (8.0.0.783), The MathWorks, Inc
38. Mather, P.M.: Computer Processing of Remotely-Sensed Images: An Introduction, 3rd edn. Wiley, New York (2004)
39. Congalton, R.G.: A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. **37**, 35–46 (1991)
40. Weber, G.-W., Batmaz, İ., Köksal, G., Taylan, P., Yerlikaya-Özkurt, F.: CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. Inverse Probl. Sci. Eng. **20**, 371–400 (2011)
41. Özmen, A., Weber, G.-W., Batmaz, İ., Kropat, E.: RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set. Commun. Nonlinear Sci. Numer. Simul. **16**, 4780–4787 (2011)
42. Pham, H.M., Yamaguchi, Y., Bui, T.Q.: A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. Landsc. Urban Plan. **100**, 223–230 (2011)
43. Zhang, Q., Wang, J., Peng, X., Gong, P., Shi, P.: Urban built-up land change detection with road density and spectral information from multi-temporal Landsat TM data. Int. J. Remote Sens. **23**, 3057–3078 (2002)
44. Kong, F., Yin, H., Nakagoshi, N., Zong, Y.: Urban green space network development for biodiversity conservation: identification based on graph theory and gravity modeling. Landsc. Urban Plan. **95**, 16–27 (2010)

45. Siart, C., Eitel, B., Panagiotopoulos, D.: Investigation of past archaeological landscapes using remote sensing and GIS: a multi-method case study from Mount Ida, Crete. J. Archaeol. Sci. **35**, 2918–2926 (2008)
46. Keay, S.J., Parcak, S.H., Strutt, K.D.: High resolution space and ground-based remote sensing and implications for landscape archaeology: the case from Portus, Italy. J. Archaeol. Sci. **52**, 277–292 (2014)
47. De Laet, V., van Loon, G.J.M., Van der Perre, A., Deliever, I., Willems, H.: Integrated remote sensing investigations of ancient quarries and road systems in the Greater Dayr al-Barshã Region, Middle Egypt: a Study of logistics. J. Archaeol. Sci. **55**, 286–300 (2015)

Dynamic Structural Econometric Modeling of the Ethanol Industry

C.-Y. Cynthia Lin Lawell

Abstract This chapter reviews some of the papers my co-authors and I have written developing and estimating dynamic structural econometric models of dynamic games in the ethanol industry. These structural econometric models model the dynamic and strategic decisions made by ethanol firms and enable us to analyze the effects of government policy. Analyses that ignore the dynamic implications of government policies, including their effects on incumbent ethanol firms' investment, production, and exit decisions and on potential entrants' entry behavior, may generate incomplete estimates of the impact of the policies and misleading predictions of the future evolution of the fuel ethanol industry. In Thome and Lin Lawell [50], we estimate a model of the investment timing game in corn ethanol plants in the United States. In Yi and Lin Lawell [53, 54], we estimate a model of the investment timing game in ethanol plants worldwide that allows for the choice among different feedstocks. In Yi et al. [55], we estimate a structural econometric model of ethanol firms' investment, production, entry, and exit decisions in order to analyze the effects of government subsidies and the Renewable Fuel Standard on the U.S. fuel ethanol industry. The results of our research will help determine which policies and factors can promote fuel-ethanol industry development.

Keywords Ethanol industry · Dynamic structural econometric model · Dynamic game

1 Introduction

Recently the support of biofuel production has been a politically sensitive topic. Politicians have pushed for support for fuel ethanol production as an environmentally friendly alternative to imported oil, as well as a way to boost farm profits and improve rural livelihoods. Several government policies actively promote ethanol production

C.-Y. C. Lin Lawell (✉)
Cornell University, Ithaca, NY, USA
e-mail: clinlawell@cornell.edu

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_14

via tax incentives and mandates, and these policies are blamed for rising food prices around the world [38]. It is important to understand the factors that have motivated the significant local investments in the ethanol industry that have been made since the mid-1990s both in the U.S. and worldwide, and, in particular, the effects of government policy.

Fuel ethanol has been in use in the United States since the time of the Model T Ford (the original flex-fuel vehicle), and while the United States passed Brazil in ethanol production in 2005, today ethanol is mostly relegated to status as a gasoline additive. The first US ethanol boom began as a result of the oil embargoes in 1973 and 1979. The desire for more energy self-sufficiency, the resulting legislation (in the form of federal income tax credits and blender's credits that continue today), and the phase out of leaded gasoline led to the construction of 153 new plants by 1985 [15]. These plants were tiny by today's standards, with an average capacity of 8 million gallons per year, and by 1991 only 35 were still operational due to poor business judgment and bad engineering [15, 51].

The second US ethanol boom began in the mid-1990s and hit full-stride by the early 2000s. Several factors contributed to this most recent boom. The Clean Air Act of 1990 mandated use of oxygenates in gasoline, of which ethanol is one, and the subsequent phase out and ban of MTBE as additive beginning in the late 1990s further increased demand for ethanol. Additionally the Renewable Fuel Standard of the Energy Policy Act of 2005 mandated ethanol production floors beginning in 2007, which rise to 36 billion gallons per year in 2033. Over this time period, the number of ethanol plants rose from 35 plants in 1991, to 50 in 1999, to 192 in September of 2010 for a total capacity of 13 billion gallons per year.

In addition to the policy and demand-side contributors to the recent ethanol boom, this new industry growth has been accompanied by changes in technology. Most significantly, the average capacity of plants in our focus region was 62 million gallons per year in 2008 up from 8 million gallons per year in 1985. In the mid-1990s the industry began designing more efficient plants, which use natural gas instead of coal as fuel [15].

In our analysis of the US ethanol industry, we focus on the second US ethanol boom. Most ethanol plants use corn as a feedstock, and thus are located in the Midwestern United States, where the majority of the corn in the US is grown. Since biofuels have been touted as a way to enhance profits in rural areas, where grain prices have remained stagnant over time, it is important to determine what factors affect decisions about when and where to invest in building new ethanol plants. In Thome and Lin Lawell [50], we model this decision using both reduced-form and structural models. In Yi et al. [55], we estimate a structural econometric model of ethanol firms' investment, production, entry, and exit decisions in order to analyze the effects of government subsidies and the Renewable Fuel Standard on the U.S. fuel ethanol industry.

Even when excluding the U.S., which was the country with the largest fuel-ethanol production in 2009, the fuel-ethanol industry has been growing rapidly in the rest of the world (ROW). Ethanol-producing countries in the ROW include Brazil, Canada, China, and Thailand, as well as countries in Europe. There are approximately 200 fuel-ethanol plants in the ROW, which is a little more than in the U.S., and over 80% of them were built after 2005. In Yi and Lin Lawell [53, 54], we estimate a model of the investment timing game in ethanol plants worldwide that allows for the choice among different feedstocks.

In Europe, 20 countries have fuel ethanol production and most of the fuel ethanol plants were built after 2000. The development of European biofuel is based on two Directives: the Renewable Energy Directive (RED) of 2003/30/EC sets indicative targets of 2% renewable fuels in transport by 2005 and 5.75% by 2010 but is not legally binding; and the RED of 2009/28/EC is made mandatory and therefore legally binding. The main fuel ethanol policies in Europe include a tax credit, a blending mandate and R&D support. Most of the policies were implemented after 2003. Empirical research shows that the effects of policies for the U.S. fuel ethanol production are positive [30, 46, 50], however, whether the stimulation effects of the government policies play the same role in Europe is not yet clear, especially for the different varieties of feedstocks. In Yi and Lin Lawell [54], we evaluate the effects of government policies on investment in ethanol plants in Europe.

The decision to invest in building an ethanol plant is a dynamic decision that may be affected by economic factors and government policies. For example, commodity markets occasionally exhibit broadly based massive booms and busts; at the core of these cycles is a set of contemporaneous supply and demand surprises that coincide with low inventories and that are magnified by macroeconomic shocks and policy responses [9]. Market volatility can induce periods of boom and bust in the ethanol industry, causing episodes of bankruptcy and reduced capital investment [23].

Because the payoff from investing in building a new ethanol plant depend on market conditions such as the feedstock price that vary stochastically over time, a potential entrant that hopes to make a dynamically optimal decision would need to account for the option value to waiting before making this irreversible investment [14].

A potential investor's investment decision may also depend on the investment decisions of other investors. When the decision of a potential investor is affected by the decisions of other investors, the decision-making problem is no longer a single-agent dynamic optimization problem, but instead becomes a multi-agent investment timing game.

There are two sources of strategic interactions that add a strategic (or non-cooperative) dimension to the potential entrants' investment timing decisions. The first source of strategic interaction is a competition effect: if there is more than one ethanol plant located in the same region, these plants may compete in the local feedstock input supply market or they may compete in the local fuel ethanol output market. The competition effect, whereby nearby plants may compete in local feedstock markets and/or local ethanol markets, deters ethanol plants from entering in regions where there are other ethanol plants already present.

The second source of strategic interaction is an agglomeration effect: if there are several ethanol plants located in the same region, the existing plants may have developed transportation and marketing infrastructure and/or an educated work force that new plants can benefit from [18, 20, 30]. The agglomeration effect induces an ethanol plant to locate near other plants, since a fuel ethanol plant benefits from the existence of other plants.

Owing to both competition and agglomeration effects, the dynamic decision-making problem faced by the potential ethanol plants is not merely a single-agent problem, but rather can be viewed as a non-cooperative game in which plants behave strategically and base decisions on other investors' strategies. Since the investment decisions of others affect future values of state variables which affect the future payoff from investing, potential investors must anticipate the investment strategies of others in order to make a dynamically optimal decision. Uncertainty over whether a plant might be constructed and start production nearby is therefore another reason there is an option value to waiting before investing [14].

In addition to the decision to build a fuel ethanol plant, ethanol firms make other decisions as well, including decisions about capacity investment, production, entry, and exit. These decisions are dynamic and strategic as well. Analyses that ignore the dynamic implications of government ethanol policies, including their effects on incumbent ethanol firms' investment, production, and exit decisions and on potential entrants' entry behavior, may generate incomplete estimates of the impact of the policies and misleading predictions of the future evolution of the fuel ethanol industry.

This article reviews some of the papers my co-authors and I have written developing and estimating dynamic structural econometric models of dynamic games in the ethanol industry. These structural econometric models model the dynamic and strategic decisions made by ethanol firms. In Thome and Lin Lawell [50], we estimate a model of the investment timing game in corn ethanol plants in the United States. This model follows my previous work estimating a structural econometric model of the multi-stage dynamic investment timing game in offshore petroleum production [35]. In Yi and Lin Lawell [53, 54], we estimate a model of the investment timing game in ethanol plants worldwide that allows for the choice among different feedstocks. In Yi et al. [55], we estimate a structural econometric model of ethanol firms' investment, production, entry, and exit decisions in order to analyze the effects of government subsidies and the Renewable Fuel Standard on the U.S. fuel ethanol industry.

Our structural econometric models enable us to model ethanol firms' strategic and dynamic investment, production, entry, and exit decisions. These decisions are dynamic because they are involve irreversible investments, because their payoffs are uncertain, and because ethanol firms have leeway over the timing of these investment decisions. Because the profits from these decisions depend on market conditions such as the ethanol and feedstock prices that vary stochastically over time, an individual firm operating in isolation that hopes to make dynamically optimal decisions would need to account for the option value to waiting before making these irreversible investments [14].

The decisions of ethanol firms are not only dynamic but strategic as well. Ethanol producers consider not only future market conditions but also their competitors' investment, production, entry and exit activities when making their current decisions. Since the production decisions of other firms affect the ethanol price, and therefore affect a firm's current payoff from production, and since the investment, production, entry, and exit decisions of other firms affect future values of state variables which affect a firm's future payoff from producing and investing, ethanol firms must anticipate the strategies of other firms in order to make a dynamically optimal decision. Uncertainty over the strategies of other firms is therefore another reason there is an option value to waiting before investing [14].

The methodology we use is to develop and estimate a structural econometric model of the dynamic game among ethanol firms. As explained by Reiss and Wolak [42], a structural econometric model is one that combines economic theory with a statistical model, enabling us to estimate structural parameters. Incorporating firm dynamics into structural econometric models enhances our understanding of behavior and also enables us to estimate structural parameters which have a transparent interpretation within the theoretical model that frames the empirical investigation [4].

Dynamic discrete choice structural models are useful tools in the analysis of economic and social phenomena whenever strategic interactions are an important aspect of individual behavior. In the ethanol market, because a firm's costs and market demand hinge on the structure of market, a firm's decision depends on its conjecture about competitors' behavior. This type of model assumes agents are forward looking and maximize the expected discounted value of the entire stream of payoffs. Agents are assumed to make decisions based only on historic information directly related to current payoffs, and history only influences current decisions insofar as it impacts a state variable that summarize the direct influence of the past on current payoffs.

There are several advantages to using a dynamic structural model to analyze the investment, production, entry, and exit decisions of ethanol firms. First, unlike reduced-form models, a structural approach explicitly models the dynamics of these decisions. Second, our dynamic games model models the strategic nature of ethanol firms' decisions.

A third advantage of the structural model is that with the structural model we are able to estimate the effect of each state variable on the expected payoffs from investment, production, entry, and exit decisions, and are therefore able to estimate parameters that have direct economic interpretations. Our dynamic model accounts for the continuation value, which is the expected value of the value function next period. With the structural model we are able to estimate parameters in the payoffs from ethanol firms' decisions, since we are able to structurally model how the continuation values relate to the payoffs from these decisions.

A fourth advantage of our structural model is that we can use the parameter estimates from our structural model to simulate various counterfactual scenarios. We use our estimates to simulate the ethanol industry under counterfactual scenarios for government policy in order to evaluate the effects of government policy.

The results of our research will help determine which policies and factors can promote fuel-ethanol industry development.

2 Literature Review

The research reviewed in this chapter builds on previous research my co-authors and I have pursued on designing and analyzing policies for renewable fuels [32, 34]; on the implications of an E10 ethanol-blend policy [31]; on the design and economics of low carbon fuel standards [27]; on the economics of California's low carbon fuel standard [24]on containing the costs of California low carbon fuel standard [24, 25, 33]; on the design of renewable fuel policies and cost containment mechanisms [29]; on ex post costs and the compliance credit market under the Renewable Fuel Standard [28]; on the effects of policy shocks that reduced the expected Renewable Fuel Standard mandates [29]; on the factors that affect ethanol investment decisions in Thailand [22]; and on the effects of China's biofuel policies on agricultural and ethanol markets [49].

Our structural model of the effects of ethanol price, corn price, gasoline price, and ethanol policy on the ethanol industry relates to the work of de Gorter et al. [12], who combine theory and empirical evidence on how biofuel policies create a link between crop (food grains and oilseeds) and biofuel (ethanol and biodiesel) prices; and on the previous literature on the relationship between food and fuel markets [1–3, 11, 40, 41, 43, 52, 56]. De Gorter et al. [13] analyze the impact of OECD biofuels policies on grain and oilseed prices in developing countries.

The dynamic structural model we will develop will build upon the dynamic structural models my co-authors and I have developed to analyze wind turbine owners' decisions about scrapping or replacing their turbines and the effects of government policies on these decisions [10]; investment decisions in offshore petroleum production [31, 35]; long-term and short-term decision-making for disease control [8]; and externalities between spinach seed companies and farmers [7].

The structural econometric models of dynamic games we use build on a model developed by Pakes et al. [39], which has been applied to the multi-stage investment timing game in offshore petroleum production [35], and to the decision to wear and use glasses [37]; a model developed by Bajari et al. [6]; as well as on a model developed by Bajari at al. [5], which has been applied to the cement industry [19, 44].

3 Theoretical Model

We model the decisions of two types of agents: incumbent ethanol plants and potential entrants in the ethanol market. Incumbents choose how much to produce, whether to invest in capacity and if so by how much, and whether to exit. Potential entrants choose whether to construct a new plant, buy a shut-down plant, or not to enter. The strategy of each agent i is assumed to be a function of a set of state variables and private information:

$$a_i = \sigma_i(s, \varepsilon_i), \tag{1}$$

where ε_i is the shock to agent i, which is not observed by either other agents or the econometrician, and where s are publicly observable state variables. State variables include own capacity, competitors' capacity, number of shut-down plants, ethanol price, feedstock price, and fuel ethanol policies.

We assume that fuel ethanol plants compete in quantities in a homogeneous goods market. The demand of fuel ethanol is homogenous over all the states, and each plant faces the national elasticity of demand. Therefore, the nation-wide fuel ethanol demand curve is given by:

$$\ln Q = \alpha_0 + \alpha_1 \ln P, \tag{2}$$

where Q is the aggregate demand for ethanol, P is the market price and α_1 is the price elasticity of demand.

For each ethanol plant i, the cost of output is assumed to be the following quadratic function of output:

$$c_i(q_i; \delta_1, \delta_2) = \delta_1 q_i + \delta_2 q_i^2, \tag{3}$$

where δ_1 and δ_2 are variable cost coefficients and q_i is the output of plant i.

Firms can change their capacities by x_i, and we assume the cost associated with capacity change is given by:

$$\Gamma(x_i; \gamma) = 1(x_i > 0)(\gamma_{1i} + \gamma_2 x_i + \gamma_3 x_i^2). \tag{4}$$

The capacity adjustment cost function shows that investment in capacity will have fixed cost γ_{1i} and quadratic variable cost with parameters γ_2 and γ_3. The individual-specific fixed cost γ_{1i}, which is private information and drawn from the distribution F_{γ_1} with mean μ_{γ_1} and standard deviation σ_{γ_1}, captures the necessary setup costs such as the costs of obtaining permits and constructing support facilities, which accrue regardless of the size of the capacity.

An ethanol plant i also faces a fixed cost $\Phi_i(a)$ unrelated to production given by:

$$\Phi_i(a_i; k, d) = \begin{cases} k_{1i} & \text{if the new entrant constructs a plant} \\ k_{2i} & \text{if the new entrant bought a plant from a previous owner} \\ -d_i & \text{if the firm exit the market} \end{cases},$$

where a_i represents the entry and exit decisions, and k_{1i} and k_{2i} are the sunk costs of entry. k_{1i} is the sunk cost of constructing a new fuel ethanol plant. Instead of constructing a new plant, another way to enter the market is to buy an existing ethanol plant that has shut down. Therefore, k_{2i} is the sunk cost of buying a shut-down plant. These sunk costs are private information and drawn from the distributions F_{k_1} and F_{k_2}, with means μ_{k_1} and μ_{k_2} and standard deviations σ_{k_1} and σ_{k_2}, respectively. If a plant exits the market, it can receive a scrap value d_i, for example from selling off the land or facility, which is private information and drawn from the distribution F_d with mean μ_d and standard deviation σ_d.

The production subsidy a fuel ethanol plant receives is:

$$r_i(q_i; \varphi) = \varphi q_i, \tag{5}$$

where φ is the subsidy level per unit of fuel ethanol.

The profit function from production for an incumbent is thus given by:

$$\bar{\pi}_i(s; \alpha, \delta_1, \delta_2, \varphi) = Pq_i - \delta_1 q_i - \delta_2 q_i^2 + \varphi q_i. \tag{6}$$

The per-period payoff function is therefore as follows:

$$\pi_i(s, a, x; \alpha, \delta, \varphi, \gamma, k, d) = \pi_i(s, a; \theta) = \bar{\pi}_i(s; \alpha, \delta, \varphi) - \Gamma(x_i; \gamma) - \Phi_i(a_i; k, d). \tag{7}$$

Hence, the value function for an incumbent, who chooses how much to produce, whether to invest in capacity and if so by how much, and whether to exit, can be represented by:

$$\begin{aligned}
&V_i(s; \sigma(s), \theta, \varepsilon_i) = \bar{\pi}_i(s; \theta) + \\
&\max\Big\{\max_{x_i > 0}\big[-\gamma_{1i} - \gamma_2 x_i - \gamma_3 x_i^2 + \beta \textstyle\int E_{\varepsilon_i'} V_i(s'; \sigma(s'), \theta, \varepsilon_i') dp(s'; s, a_i, \sigma_{-i}(s))\big], \\
&\beta \textstyle\int E_{\varepsilon_i'} V_i(s'; \sigma(s'), \theta, \varepsilon_i') dp(s'; s, a_i, \sigma_{-i}(s)),\ d_i\Big\},
\end{aligned}$$

where the continuation value $\int E_{\varepsilon_i'} V_i(s'; \sigma(s'), \theta, \varepsilon_i') dp(s'; s, \sigma(s))$ is the expected value of the value function next period conditional on the state variables and strategies in the current period, s' is the vector of next period's state variables, $p(s'; s, a_i, \sigma_{-i}(s))$ is the conditional probability of state variable s' given the current state s, player i's action a_i (including any capacity changes x_i) and the strategies $\sigma_{-i}(s)$ of all other players. Incumbents receive the profits $\bar{\pi}_i(s; \theta)$ from production this period and then, depending on their action, additionally incur the costs of capacity investment if they invest, additionally receive the continuation value if they stay in the market (regardless of whether they invest), and additionally receive the scrap value from exiting if they exit.

Similarly, the value function for a potential entrant, who can either stay out of the ethanol market, build a new plant or buy a shut-down plant from a previous owner, is:

$$\begin{aligned}
&V_i(s; \sigma(s), \theta, \varepsilon_i) = \max\Big\{\varepsilon_{0i}, \\
&\max_{y_i > 0}\Big[-k_{1i} - \gamma_{1i} - \gamma_2 y_i - \gamma_3 y_i^2 + \varepsilon_{1i} + \beta \textstyle\int E_{\varepsilon_i} V_i(s'; \sigma(s', \theta, \varepsilon_i)) dp(s'; s, a_i, \sigma_{-i}(s))\Big], \\
&\max_{\substack{y_i > 0, \\ y_i \in \boldsymbol{Y}}}\Big[-k_{2i} - \gamma_4 y_i - \gamma_5 y_i^2 + \varepsilon_{2i} + \beta \textstyle\int E_{\varepsilon_i} V_i(s'; \sigma(s', \theta, \varepsilon_i)) dp(s'; s, a_i, \sigma_{-i}(s))\Big]\Big\},
\end{aligned}$$

where y_i is the capacity for plant i; γ_4 and γ_5 are transaction cost parameters for an entrant buying an shut-down plant; $\boldsymbol{Y}$ is the set of shut-down plants' sizes in the market; and ε_{0i}, ε_{1i} and ε_{2i} are idiosyncratic preference shocks that we assume are independently distributed with an extreme value distribution. The value function for a potential entrant is therefore the maximum of: (1) the payoff from staying out of the market, which is the idiosyncratic preference shock ε_{0i}; (2) the payoff from building a new plant, which includes the fixed cost of entry k_{1i}, the costs of capacity investment, the idiosyncratic preference shock ε_{1i}, and the continuation value; and (3) the payoff from building a shut-down plant, which includes the fixed cost of entry k_{2i}, the transactions costs, the idiosyncratic preference shock ε_{2i}, and the continuation value. If an entrant decides to buy an existing shut-down plant, its plant size choice is limited to set $\boldsymbol{Y}$.

We assume that each plant optimizes its behavior conditional on the current state variables, other agents' strategies and its own private shocks, which results in a Markov perfect equilibrium (MPE). The optimal strategy $\sigma_i^*(s)$ for each player i should therefore satisfy the following condition for all state variables s and alternative strategies $\tilde{\sigma}_i(s)$:

$$V_i(s; \sigma_i^*(s), \sigma_{-i}, \theta, \varepsilon_i) \geq V_i(s; \tilde{\sigma}_i(s), \sigma_{-i}, \theta, \varepsilon_i).$$

4 Econometric Methodology

In Thome and Lin Lawell [50], we estimate a model of the investment timing game in corn ethanol plants in the United States. This model follows my previous work estimating a structural econometric model of the multi-stage dynamic investment timing game in offshore petroleum production [35], which is based on an econometric model developed by Pakes et al. [39]. In Lin [35], I build on the work of Pakes et al. [39] on discrete games of entry and exit by examining sequential investments with a finite horizon. The econometric estimation technique takes place in two steps. In the first step, the continuation value is estimated nonparametrically and used to form the model's estimate of the investment probabilities. In the second step, the investment probabilities predicted by the model are matched with the empirical investment probabilities in the data using generalized method of moments.

In Yi and Lin Lawell [53, 54], we estimate a model of the investment timing game in ethanol plants worldwide that allows for the choice among different feedstocks. We use a structural model developed by Bajari et al. [6]. We construct a dynamic discrete choice model for a potential fuel ethanol plant in which the investor maximizes its present discounted value of its entire stream of payoffs, and in which the decisions of other plants in the same local market affect an investor's decision. The innovative features of our model are the consideration of interactions between fuel ethanol plants and the dynamic decision making framework. The effects of economic, policy and strategic variables on per-period profit are estimated via a semiparametric approach.

Our research in Yi and Lin Lawell [53, 54] differs from previous studies of the investment and location of ethanol plants because it models the decision as a dynamic one rather than a static one, because it allows for the choice among multiple feedstocks rather than just one feedstock such as corn, because its strategic framework allows the estimation of strategic interactions among plants, and because it uses international data rather than data from the U.S.

In Yi, Lin Lawell and Thome [55], we analyze how government subsidies and the renewable fuels standard affect fuel ethanol production, investment, entry, and exit by estimating a structural econometric model of a dynamic game. We use the structural econometric model developed by Bajari et al. [5] and applied by Ryan [44] to evaluate the effects of environmental regulation on the U.S. cement industry. In particular, we assume that each plant optimizes its behavior conditional on the current state variables including other agents' actions and its own private shocks, which results in a Markov perfect equilibrium (MPE). We estimate the structural econometric model in two steps. In the first step, we characterize the policy functions for the plants' decisions regarding entry, capacity investment and exit, which are functions of state variables. In the second step, we use a simulation-based minimum distance estimator proposed by Bajari et al. [5] to estimate the distribution of fixed costs and the variable costs for changing ethanol plant capacity, the distribution of scrap values a plant would receive if it exited the market, and the distribution of entry costs and the variable costs for either constructing a new plant or buying a shut-down plant.

In Yi et al. [55], we build upon the previous literature by estimating the various investment and production costs empirically, and also by allowing for two different types of entry: entry via constructing a new plant and entry via buying a shut-down plant. An additional innovation in our paper is that we allow our estimated cost parameters to depend on production subsidy levels and on the implementation of the Renewable Fuel Standard. In contrast to our paper, which empirically estimates costs, the cost information used in previous studies of the ethanol industry are mainly from the literature or from engineering experiments [16, 17, 21, 47, 48].

5 Results

The results of our structural model in Thome and Lin Lawell [50] show that in the US, the intensity of corn production; government policies, particularly the MTBE ban and the 2007 Renewable Fuel Standard (RFS2); and private information shocks all have significant effects on ethanol investment payoffs and decisions. We use the estimated structural parameters to simulate counterfactual policy scenarios to disentangle the impacts of state and national policies on the timing and location of investment in the industry. We find that, of the policies analyzed, the MTBE ban and the RFS2 led to most of the investment during this time period.

Our results in Yi and Lin Lawell [54] show that in Europe, competition between plants deters local investments and has a large negative effect on the payoffs from investment. We also find that government policies have a large positive effect on payoffs from investment. Ethanol investment decisions in Europe are affected more by government policies and strategic interactions than by economic factors.

Our results in Yi and Lin Lawell [53] show that in Canada, competition between plants is enough to deter local investments, the availability of feedstock is important in determining plant location, and the effects of policy support for wheat-based plants are significant.

Our empirical results in Yi et al. [55] show that the production subsidy does not affect either investment costs or scrap values, but the Renewable Fuel Standard significantly impacts the distributions of both the fixed cost of plant capacity investment and the scrap value a plant would receive if it exited the market. We then use our estimated structural model of the fuel ethanol industry to simulate the effects of 3 different types of subsidy: a volumetric production subsidy, an investment subsidy, and an entry subsidy, each with and without the Renewable Fuel Standard. Results show that the Renewable Fuel Standard is a critically important policy for supporting the sustainability of corn-based fuel ethanol production. In addition, we find that investment subsidies and entry subsidies are more effective than production subsidies and that with an investment subsidy or an entry subsidy the government can pay much less than it would under a production subsidy but still reach the goal set by the Renewable Fuel Standard.

6 Conclusions

This chapter reviews some of the papers my co-authors and I have written developing and estimating dynamic structural econometric models of dynamic games in the ethanol industry. These structural econometric models model the dynamic and strategic decisions made by ethanol firms and enable us to analyze the effects of government policy. Analyses that ignore the dynamic implications of government policies, including their effects on incumbent ethanol firms' investment, production, and exit decisions and on potential entrants' entry behavior, may generate incomplete estimates of the impact of the policies and misleading predictions of the future evolution of the fuel ethanol industry.

According to our results, we find in Thome and Lin Lawell [50] that, in the United States, the intensity of corn production; government policies, particularly the MTBE ban and the 2007 Renewable Fuel Standard (RFS2); and private information shocks all have significant effects on ethanol investment payoffs and decisions. For Europe, we find in Yi and Lin Lawell [54] that competition between plants deters local investments and ethanol support policies encourage investments. For Canada, we find in Yi and Lin Lawell [53] that competition between plants is enough to deter local investments, the availability of feedstock is important in determining plant location, and the effects of policy support for wheat-based plants are significant.

Our results in Yi et al. [55] show that the Renewable Fuel Standard is a critically important policy for supporting the sustainability of corn-based fuel ethanol production. In addition, we find that investment subsidies and entry subsidies are more effective than production subsidies and that with an investment subsidy or an entry subsidy the government can pay much less than it would under a production subsidy but still reach the goal set by the Renewable Fuel Standard.

Our results have important implications for the design of government policies for ethanol. In particular, the results of our research will help determine which policies and factors can promote fuel-ethanol industry development.

References

1. Abbott, P., Hurt, C., Tyner, W.E.: Whats Driving Food Prices?. Farm Foundation Issue Report (2008)
2. Abbott, P., Hurt, C., Tyner, W.E.: What's Driving Food Prices?. March 2009 Update in Farm Foundation Issue Report 2009 (2009)
3. Abbott, P., Hurt, C., Tyner, W.E.: What's Driving Food Prices in 2011?. Farm Foundation Issue Report 2011 (2011)
4. Aguirregabiria, V., Mira, P.: Dynamic discrete choice structural models: a survey. J. Econom. **156**(1), 38–67 (2010)
5. Bajari, P., Benkard, C.L., Levin, J.: Estimating dynamic models of imperfect competition. Econometrica **75**(5), 1331–1370 (2007)
6. Bajari, P., Chernozhukov, V., Hong, H., Nekipelov, D.: Identification and efficient semiparametric estimation of a dynamic discrete game. NBER Working paper 21125 (2015)
7. Carroll, C.L., Carter, C.A., Goodhue, R.E., Lin Lawell, C.-Y.C.: Supply chain externalities and agricultural disease. Working paper (2017a)
8. Carroll, C.L., Carter, C.A., Goodhue, R.E., Lin Lawell, C.-Y.C.: The economics of decision-making for crop disease control. Working paper (2017b)
9. Carter, C.A., Rausser, G.C., Smith, A.: Commodity booms and busts. Annu. Rev. Resour. Econ. **3**, 87–118 (2011)
10. Lin Lawell, C.-Y.C.: Wind turbine shutdowns and upgrades in Denmark: Timing decisions and the impact of government policy. Working paper, Cornell University (2017)
11. de Gorter, H., Drabik, D., Just, D.R.: How biofuels policies affect the level of grains and oilseed prices: theory, models and evidence. Glob. Food Secur. **2**, 82–88 (2013)
12. de Gorter, H., Drabik, D., Just, D.R.: The Economics of Biofuel Policies: Impacts on Price Volatility in Grain and Oilseed Markets. Palgrave-McMillan, New York (2015)
13. de Gorter, H., Drabik, D., Just, D.R., Kliauga, E.M.: The impact of OECD biofuels policies on developing countries. Agric. Econ. **44**, 477–486 (2013)
14. Dixit, A.K., Pindyck, R.S.: Investment Under Uncertainty. Princeton University Press, Princeton (1994)
15. DOE [Deparment of Energy] (2008). Energy Time Lines: Ethanol. Revised June, 2008. Accessed Jan 2009
16. Eidman, V.R.: Ethanol economics of dry mill plants. Corn-Based Ethanol in Illinois and the US: A Report from the Department of Agricultural and Consumer Economics, University of Illinois, 22–36. (2007)
17. Ellinger, P.N.: Assessing the financial performance and returns of ethanol production: a case study analysis. Corn-Based Ethanol in Illinois and the US: A Report from the Department of Agricultural and Consumer Economics, University of Illinois, 38–62. (2007)
18. Ellison, G., Glaeser, E.L.: The geographic concentration of industry: does natural advantage explain agglomeration? Am. Econ. Rev. **89**(2), 311–316 (1999)
19. Fowlie, M., Reguant, M., Ryan, S.P.: Market-based emissions regulation and industry dynamics. J. Polit. Econ. **124**(1), 249–302 (2016)
20. Goetz, S.: State- and county-level determinants of food manufacturing establishment growth: 1987–93. Am. J. Agric. Econ. **79**, 838–850 (1997)
21. Gonzalez, A.O., Karali, B., Wetzstein, M.E.: A public policy aid for bioenergy investment: case study of failed plants. Energy Policy **51**, 465–473 (2012)
22. Herath Mudiyanselage, N., Lin, C.-Y.C., Yi, F.: An analysis of ethanol investment decisions in Thailand. Theor. Econ. Lett. **3**(5A1), 14–20 (2013)
23. Hochman, G., Sexton, S.E.: The economics of biofuel policy and biotechnology. J. Agric. Food Ind. Organ. **6**(2), (2008). Article 8
24. Lade, G.E., Lin, C.-Y.C.: A report on the economics of Californias low carbon fuel standard and cost containment mechanisms. Prepared for the California Air Resources Board. Institute of Transportation Studies, University of California at Davis, Research Report UCD-ITS-RR-13–23 (2013)

25. Lade, G.E., Lin, C.-Y.C.: Controlling compliance costs for Californias LCFS with a price ceiling. University of California at Davis Institute of Transportation Studies, Policy brief (2014)
26. Lade, G.E., Lin Lawell, C.-Y.C. Lin, C.-Y.C.: The design of renewable fuel policies and cost containment mechanisms. Working paper (2017)
27. Lade, G.E., Lin Lawell, C.-Y.C.: The design and economics of low carbon fuel standards. Res. Transp. Econ. **52**, 91–99 (2015)
28. Lade, G.E., Lin, C.-Y.C., Smith, A.: Ex post costs and renewable identification number (RIN) prices under the Renewable Fuel Standard. Resources for the Future Discussion Paper 15-22. (2015)
29. Lade, G.E., Lin Lawell, C.-Y.C., Smith, A.: Policy shocks and market-based regulations: Evidence from the Renewable Fuel Standard. Working paper (2017)
30. Lambert, D.M., Wilcox, M., English, A., Stewart, L.: Ethanol plant location determinants and county comparative advantage. J. Agric. Appl. Econ. **40**, 117–135 (2008)
31. Lin, C.-Y.C.: Estimating strategic interactions in petroleum exploration. Energy Econ. **31**(4), 586–594 (2009)
32. Lin, C.-Y.C.: On designing and analyzing policies for renewable fuels. California State Controller John Chiang statement of general fund cash receipts and disbursements **6**(12), 4–5 (2012)
33. Lin, C.-Y.C.: Containing the costs of California's low carbon fuel standard. California State Controller John chiang statement of general fund cash receipts and disbursements **7**(12), (2013a)
34. Lin, C.-Y.C.: On designing and analyzing policies for renewable fuels. Energy Dimensions. http://www.energydimensions.net/on-designing-and-analyzing-policies-for-renewable-fuels/ (2013b)
35. Lin, C.-Y.C.: Strategic decision-making with information and extraction externalities: a structural model of the multi-stage investment timing game in offshore petroleum production. Rev. Econ. Stat. **95**(5), 1601–1621 (2013c)
36. Lin, C.-Y.C., Zhang, W., Rouhani, O., Prince, L.: The implications of an E10 ethanol-blend policy for California. Agric. Resour. Econ. Update **13**(2), 1–4 (2009)
37. Ma, X., Lin Lawell, C.-Y.C., Rozelle, S.: Estimating peer effects: a structural econometric model using a field experiment of a health promotion program in rural China. Working paper, Cornell University (2017)
38. Mitchell, D.: A note on rising food prices. Policy research Working Paper # 4862, The World Bank Development Prospects Group, July 2008 (2008)
39. Pakes, A., Ostrovsky, M., Berry, S.: Simple estimators for the parameters of discrete dynamic games (with entry and exit examples). RAND J. Econ. **38**(2), 373 (2007)
40. Poudel, B.N., Paudel, K.P., Timilsina, G., Zilberman, D.: Providing numbers for a food versus fuel debate: an analysis of a future biofuel production scenario. Appl. Econ. Perspect. Policy **34**(4), 637–668 (2012)
41. Rajagopal, D., Sexton, S., Roland-Holst, D., Zilberman, D.: Challenge of biofuel: filling the tank without emptying the stomach? Environ. Res. Lett. **2**(4), 1–9 (2007)
42. Reiss, P.C., Wolak, F.A.: In: Heckman, J.J., Leamer, E.E. (eds.) Structural Econometric Modeling: Rationales and Examples from Industrial Organization. Handbook of Econometrics, vol. 6A, pp. 4277–4415. Stanford, California (2007)
43. Runge, C.F., Senauer, B.: How Biofuels Could Starve the Poor. Foreign Affairs, 41–53. (2007)
44. Ryan, S.P.: The costs of environmental regulation in a concentrated industry. Econometrica **80**(3), 1019–1061 (2012)
45. Sarmiento, C., Wilson, W.W.: Spatial competition and ethanol plant location decisions. July 2008, 2008 Annual Meeting. American Agricultural Economics Association, Orlando, Florida 6175, (2008)
46. Sarmiento, C., Wilson W.W., Dahl, B. Spatial competition and ethanol plant location decisions. Agribusiness **28**(3), 260–273 (2012)
47. Schmit, T.M., Luo, J., Conrad, J.M.: Estimating the influence of US ethanol policy on plant investment decisions: a real options analysis with two stochastic variables. Energy Econ. **33**(6), 1194–1205 (2011)

48. Schmit, T.M., Luo, J., Tauer, L.W.: Ethanol plant investment using net present value and real options analyses. Biomass Bioenergy **33**(10), 1442–1451 (2009)
49. Si, S., Chalfant, J.A., Lin Lawell, C.-Y.C., Yi, F.: The effects of China's biofuel policies on agricultural and ethanol markets. Working paper, Cornell University (2017)
50. Thome, K.E., Lin Lawell, C.-Y.C.: Investment in corn-ethanol plants in the Midwestern United States. Working paper, Cornell University (2017)
51. Urbanchuk, J.M.: Economic Impacts on the Farm Community of Cooperative Ownership of Ethanol Production. National Corn Growers Association Report (2006)
52. Wright, B.: Global biofuels: key to the puzzle of grain market behavior. J. Econ. Persp. **28**(1), 73–98 (2014)
53. Yi, F., Lin Lawell, C.-Y.C.: Ethanol plant investment in Canada: a structural model. Working paper, Cornell University (2017a)
54. Yi, F., Lin Lawell, C.-Y.C.: What factors affect the decision to invest in a fuel ethanol plant?: a structural model of the ethanol investment timing game. Working paper, Cornell University (2017b)
55. Yi, F., Lin Lawell, C.-Y.C., Thome, K.E.: The effects of subsidies and mandates: a dynamic model of the ethanol industry. Working paper, Cornell University (2017)
56. Zilberman, D., Hochman, G., Rajagopal, D., Sexton, S., Timilsina, G.: The impact of biofuels on commodity food prices: assessment of findings. Am. J. Agric. Econ. **95**, 275–281 (2012)

An Introduction to Coupling

Artur O. Lopes

Abstract In this review paper we describe the use of couplings in several different mathematical problems. We consider the total variation norm, maximal coupling and the $\bar{d}$-distance. We present a detailed proof of a result recently proved: the dual of the Ruelle operator is a contraction with respect to 1-Wasserstein distance. We also show the exponential convergence to equilibrium on the state space for finite state Markov chains when the transition matrix $\mathscr{P}$ has all entries positive.

Keywords Coupling · Total variation norm · Maximal coupling · Transport · Optimal plan · Wasserstein distance · Exponential convergence to equilibrium · Dual of the Ruelle operator

1 Introduction

This is a review paper presenting some examples which were described in the literature where couplings are used for deriving results in Ergodic Theory and Probability. We present several simple calculations which we believe can help the newcomer to the subject. We are writing for a broad audience and not for the expert.

Our purpose is to present such results in a more direct approach. This will avoid the reader which is interested in the topic to have to look in several different references.

We describe the results in a language which is more familiar to the Dynamical Systems audience.

In Sect. 2 we present some definitions and mention some related results which are required in Sect. 7 where we consider the dual of the Ruelle operator.

In Sects. 3 and 4 we consider the total variation norm and the maximal coupling (see Theorem 1).

In Sect. 6 we consider the d-bar distance among Bernoulli probabilities (see Theorem 3).

A.O. Lopes (✉)
Instituto de Matemática-UFRGS, Avenida Bento Gonçalves 9500, Porto Alegre-RS, Brazil
e-mail: alopes@mat.ufrgs.br

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_15

We also show the exponential convergence to equilibrium on the state space for finite state Markov chains (see Theorem 2) when the transition matrix $\mathscr{P}$ has all entries positive (see Sects. 5.1 and 5.2).

The use of coupling and the Wasserstein distance can be useful for estimating decay of correlations (see [3, 9, 21]) and also in other different dynamical and ergodic problems (see [1, 13]).

We believe is important to describe interesting plans or couplings that can be used in estimations of different nature. Variations of these plans can be helpful to solve other open problems.

According to den Hollander [5]: *coupling is an art, not a recipe*.

We would like to thanks Anthony Quas, Diogo Gomes, Adriana Neumann and Rafael Souza for helpful conversations during the period we were writing this review paper.

2 Some Definitions and the Dual of the Ruelle Operator

Definition 1 Given the Bernoulli space $\Omega = \{1, 2, ..., d\}^{\mathbb{N}}$ each Γ in the set of probabilities on $\Omega \times \Omega$ is called a plan.

Given the Bernoulli space $\Omega = \{1, 2, ..., d\}^{\mathbb{N}}$ and two probabilities μ and ν on the natural Borel sigma algebra of Ω, a coupling of μ and ν is a plan Γ on the product space $\Omega \times \Omega$, such that, the first marginal of Γ is μ and the second is ν.

$\mathscr{C}(\mu, \nu)$ by definition is the set of plans Γ on $\Omega \times \Omega$ such that the projection in the first coordinate is μ and in the second is ν.

Definition 2 Given a distance d on Ω we denote

$$W_1(\mu, \nu) = \inf_{\Gamma \in \mathscr{C}(\mu,\nu)} \int d(x, y) d\,\Gamma(dx, dy).$$

The above expression defines a metric on the space of probabilities over Ω which is compatible with the weak$*$-convergence (see [23]). This value is called the 1-Wasserstein distance of μ and ν.

Each plan (it can exist more than one) which realizes the above infimum is called an optimal plan for d. We denote by $d_1(\mu, \nu) = d_1(\mu, \nu)$ the corresponding metric in the set of probabilities on Ω.

Definition 3 More generally, given a continuous function $c : \Omega \times \Omega \to \mathbb{R}$ and fixed μ and ν, one can be interested in

$$\inf_{\Gamma \in \mathscr{C}(\mu,\nu)} \int c(x, y) d\,\Gamma(dx, dy).$$

Each plan Γ (it can exist more than one) which realizes the above infimum is called an optimal plan, or optimal coupling, for c and the pair μ and ν.

In general is not so easy to identify exactly the optimal plan Γ. Anyway, if we are lucky to find some plan which is almost optimal then we can get some interesting results. In simple words this is the main issue on Coupling Theory.

The Kantorovich duality theorem (see [23]) is a main result which also helps to get good estimates in problems of different nature.

The total variation norm (to be defined later) is a different way to measure the distance among two probabilities. It is not equivalent to weak-$*$-convergence. This will be also considered in the text.

One of our purposes is to illustrate through several examples the use of coupling in interesting problems.

Suppose $\theta < 1$ is fixed. On the Bernoulli space $\Omega = \{1, 2, ..., d\}^{\mathbb{N}}$ we consider the metric d_θ. By definition $d_\theta(x, y) = \theta^N$ where $x_1 = y_1, .., x_{N-1} = y_{N-1}$ and $x_N \neq y_N$.

We briefly mention some properties related to Gibbs states of Lipchitz potentials (see [18] for general results).

Definition 4 Given $A : \Omega \to \mathbb{R}$, the Ruelle operator $\mathscr{L}_A$ acts on functions $\psi : \Omega \to \mathbb{R}$ in the following way

$$\varphi(x) = \mathscr{L}_A(\psi)(x) = \sum_{a=1}^{d} e^{A(ax)} \psi(ax).$$

By this we mean $\mathscr{L}_A(\psi) = \varphi$.

Suppose $A = \log J$ is Lipchitz and normalized, that is, for any $x \in \Omega$ we have $\mathscr{L}_{\log J}(1)(x) = 1$.

All probabilities we consider will be over the Borel sigma-algebra $\mathscr{B}$ of Ω.

Definition 5 Given the continuous potential $\log J : \Omega \to \mathbb{R}$ let $\mathscr{L}^*_{\log J}$ be the operator on the set of Borel Measures on Ω defined so that $\mathscr{L}^*_{\log J}(\nu)$, for each Borel measure ν, satisfies:

$$\int_\Omega \psi \, d\mathscr{L}^*_{\log J}(\nu) = \int_\Omega \mathscr{L}_{\log J}(\psi) \, d\nu,$$

for all continuous functions ψ.

$\mathscr{L}^*_{\log J}$ takes probabilities in probabilities.

A probability which is fixed for such operator $\mathscr{L}^*_{\log J}$ is invariant for the shift and called a g-measure. In the case $\log J$ is Holder such fixed point is unique and is the Gibbs state for $\log J$ (see [18]).

Is it true that there exist a metric d equivalent to d_θ such that $\mathscr{L}^*_{\log J}$ is a contraction in the 1 Wassertein distance d_1 associated to such d? The answer to this question is yes and we will address the question in the Sect. 7. Before that we will present some more basic material in the next sections.

Stadlbauer presented a proof with an affirmative answer to the above question in a more general setting. The present proof is just a simplified version of the one in [20]. The proof is based in an adaptation to our case of the more general setting described in [12].

This is true indeed. A proof of this fact which applies for the Thermodynamic Formalism setting (the one above) and also for some iterated contraction systems appears in [14].

Stadlbauer presented a proof with an affirmative answer to the above question in a more general setting [20].

3 The Total Variation Norm and the Maximal Coupling

Suppose ρ is a signed measure in the Borel sigma-algebra $\mathscr{B}$ of $\Omega = \{1, 2, .., d\}^{\mathbb{N}}$ such that $\rho(\Omega) = 0$.

Definition 6 The total variation of a signed measure ρ as above is by definition

$$|\rho|_{tv} = 2 \sup_{A \in \mathscr{B}} \rho(A).$$

One can show by duality that

$$|\rho|_{tv} = \sup_{|\phi|_\infty \leq 1} \int \phi \, d\rho.$$

where ϕ are measurable and bounded and $|\phi|_\infty$ is the supremum norm.

Given two probabilites μ_1 and μ_2 in Ω one can consider the distance $|\mu_1 - \mu_2|_{tv}$, called the total variation distance of μ_1 and μ_2. This is a different way to measure the distance among two probabilities.

This distance is also known as the strong distance in opposition to the more well known concept of weak convergence of probabilities.

Remember that we denote $\mathscr{C}(\mu_1, \mu_2)$ the set plans Γ in $\Omega \times \Omega$ such that the projection in the first coordinate is μ_1 and in the second is μ_2.

Proposition 1 **:** *Given $\Gamma \in \mathscr{C}(\mu_1, \mu_2)$, then,*

$$|\mu_1 - \mu_2|_{tv} \leq 2 \; \Gamma \, \{(x, y) \,|\, x \neq y\}.$$

Proof Given $A \in \mathscr{B}$ we have that

$$\mu_1(A) - \mu_2(A) = \int \int_{x \in A} \Gamma(dx, dy) - \int_{y \in A} \int \Gamma(dx, dy) =$$

$$[\int_{x=y}\int_{x\in A} \Gamma(dx,dy) + \int_{x\neq y}\int_{x\in A} \Gamma(dx,dy)\,]-$$

$$[\int_{y\in A}\int_{x=y} \Gamma(dx,dy) + \int_{y\in A}\int_{x\neq y} \Gamma(dx,dy)\,] =$$

$$\int_{x\neq y}\int_{x\in A} \Gamma(dx,dy) - \int_{y\in A}\int_{x\neq y} \Gamma(dx,dy) \leq \int_{x\neq y}\int_{x\in A} \Gamma(dx,dy).$$

Remember that

$$|\mu_1 - \mu_2|_{tv} = 2 \sup_{A\in\mathscr{B}} (\mu_1 - \mu_2)(A).$$

Taking supremum in $A \in \mathscr{B}$ we get the claim. □

We follow the general reasoning of [5, 16].

Suppose μ_1 and μ_2 are two probabilities on the Bernoulli space Ω.

Put $\lambda = \mu_1 + \mu_2$, and

$$g = \frac{d\mu_1}{d\lambda},\; g' = \frac{d\mu_2}{d\lambda}.$$

Now we define Q on Ω by $d\,Q = (\,g \wedge g'\,)\,d\lambda$, where $g \wedge g'$ denotes the infimum of g and g'.

Note that by Kantorovich Duality (see [23])

$$|\mu_1 - \mu_2|_{tv} = \sup_{|f|\leq 1,\ f \text{ measurable}} |\int f\,d\mu_1 - \int f\,d\mu_2\,|.$$

Therefore,

$$|\mu_1 - \mu_2|_{tv} = \sup_{|f|\leq 1,\ f \text{ measurable}} |\int f\,(g-g')d\lambda\,| =$$

$$\int_{g\geq g'} 1\,(g-g')d\lambda + \int_{g<g'} (-1)\,(g-g')d\lambda = \int |\,g-g'\,|d\lambda. \tag{1}$$

Consider $\varphi : \Omega \to \Omega \times \Omega$ by $\varphi(x) = (x,x)$.

Finally, we denote by $\hat{Q} = \varphi^*(Q)$.

Note that the support of $\hat{Q}$ is the diagonal Δ in $\Omega \times \Omega$.

Now, let $\gamma = \hat{Q}(\Delta) = Q(\Omega)$.

We call $\nu_1 = \mu_1 - Q$ and $\nu_2 = \mu_2 - Q$, and finally we define a plan π in $\Omega \times \Omega$ by

$$\pi = \hat{Q} + \frac{\nu_1 \otimes \nu_2}{1-\gamma}.$$

This plan is sometimes called **maximal coupling**.

Note that $\nu_2(\Omega) = \mu_2(\Omega) - Q(\Omega) = 1 - \gamma = \nu_1(\Omega) = \mu_1(\Omega) - Q(\Omega)$.

We claim that π projects in the first coordinate on μ_1. Indeed,

$$\pi(A \times \Omega) = \hat{Q}(A \times \Omega) + \frac{\nu_1 \otimes \nu_2}{1-\gamma}(A \times \Omega) =$$

$$\hat{Q}(A \times A) + \frac{\nu_1(A) \otimes \nu_2(\Omega)}{1-\gamma} =$$

$$Q(A) + \frac{\nu_1(A)\,(1-\gamma)}{1-\gamma} = Q(A) + (\mu_1(A) - Q(A)) = \mu_1(A).$$

The above also shows that π is a probability.

Moreover, π projects in the second coordinate on μ_2. Indeed,

$$\pi(\Omega \times A) = \hat{Q}(\Omega \times A) + \frac{\nu_1 \otimes \nu_2}{1-\gamma}(\Omega \times A) =$$

$$\hat{Q}(A \times A) + \frac{\nu_1(\Omega) \otimes \nu_2(A)}{1-\gamma} =$$

$$Q(A) + \frac{\nu_2(A)\,(1-\gamma)}{1-\gamma} = Q(A) + (\mu_2(A) - Q(A)) = \mu_2(A).$$

In this way $\pi \in \mathscr{C}(\mu_1, \mu_2)$. Therefore, it follows from a previous result that for such plan it is true the property $|\mu_1 - \mu_2|_{tv} \leq 2\,\pi\,\{(x, y)\,|\,x \neq y\}$. Now we will show:

Theorem 1 *The plan π defined above satisfies*

$$|\mu_1 - \mu_2|_{tv} = 2\,\pi\,\{(x, y)\,|\,x \neq y\}.$$

Proof First note that as $|g - g'| = g + g' - 2\,(g \,\wedge g')$, we have by (1)

$$|\mu_1 - \mu_2|_{tv} = \int |g\, - g'|\,d\,\lambda = 2\,[\,1 - \int (\,g\, \wedge g'\,)\,d\,\lambda\,] =$$

$$2\,(1 - Q(\Omega)) = 2\,(1-\gamma)\, \geq\, 2\,\pi(\Delta^c) = 2\,\pi\,\{(x, y)\,|\,x \neq y\}.$$

The last inequality follows from

$$\pi(\Delta^c) = \hat{Q}(\Delta^c) + \frac{\nu_1 \otimes \nu_2(\Delta^c)}{1-\gamma} \leq$$

$$\frac{\nu_1 \otimes \nu_2(\Delta^c)}{1-\gamma} \leq \frac{\nu_1 \otimes \nu_2(\Omega \times \Omega)}{1-\gamma} = \frac{\nu_1(\Omega) \times \nu_2(\Omega)}{1-\gamma} = \frac{(1-\gamma)^2}{1-\gamma} = 1-\gamma.$$

□

Given a probability ν on the Bernoulli space Ω then the probability $\mu = \sigma^*(\nu)$ is by definition the one such that $\mu(A) = \nu(\sigma^{-1}(A))$ for any Borel set A on Ω. We say that μ is invariant for the shift if $\mu = \sigma^*(\mu)$.

Proposition 2 *Given μ_1 and μ_2 two probabilities over Ω, then*

$$|\sigma^*(\mu_1) - \sigma^*(\mu_2)|_{tv} \leq |\mu_1 - \mu_2|_{tv}.$$

Proof Note that by Kantorovich Duality (see [23])

$$|\sigma^*(\mu_1) - \sigma^*(\mu_2)|_{tv} = \sup_{|f|\leq 1,\, f \text{ measurable}} |\int f\, d\sigma^*\mu_1 - \int f\, d\sigma^*\mu_2\,| =$$

$$\sup_{|f|\leq 1,\, f \text{ measurable}} |\int (f \circ \sigma)\, d\mu_1 - \int (f \circ \sigma)\, d\mu_2\,|.$$

The functions of the form $(f \circ \sigma)$ with $|f| \leq 1$ is a smaller class that the set of functions of the form g such that $|g| \leq 1$.

From this follows the claim. □

In this way the composition with the shift never increase the total variation norm of probabilities.

4 The Estimation Using T

Remember that points x in $\Omega = \{1, 2, ..., d\}^{\mathbb{N}}$ are denoted by $x = (x_1, x_2, x_3, ...)$.

Definition 7 The coupling time T is **the** measurable function $T : \Omega \times \Omega \to \mathbb{N}$ given by

$$T(x, y) = \inf\{n \mid x_m = y_m \text{ for all}\, m \geq n\},$$

for any $x, y \in \Omega$.

This value can be eventually ∞.

An alternative way to define the coupling time T is given by

$$T(x, y) = \inf\{n \mid \sigma^n(x) = \sigma^n(y)\},$$

for any $x, y \in \Omega$.

Note that there is a difference in estimating the total variation of two probabilities in the state space $\{1, 2, .., d\}$ and in the Bernoulli space $\{1, 2, .., d\}^{\mathbb{N}}$. We consider bellow results for each kind of setting.

We use the notation: for any n we have that X_n is the measurable function $X_n : \{1, 2, .., d\}^{\mathbb{N}} \to \{1, 2, .., d\}$ such that $X_n(x) = x_n$ if $x = (x_1, x_2, x_3, ..., x_n, ...)$.

Proposition 3 *Estimation in the state space - Suppose μ_1 and μ_2 are probabilities on $\{1, 2, .., d\}$. Suppose also that Γ is a probability on $\{1, 2, .., d\}^{\mathbb{N}} \times \{1, 2, .., d\}^{\mathbb{N}}$, such that for a fixed n,*

for all $J \subset \{1, 2..., d\}$, we have

$$\Gamma(X_n \in J, X_n \in \{1, 2, .., d\}) = \mu_1(J) \tag{2}$$

and for all $J \subset \{1, 2..., d\}$, we have

$$\Gamma(X_n \in \{1, 2, .., d\}, X_n \in J) = \mu_2(J). \tag{3}$$

Then,

$$\mu_1 \{x \,|x_n \in J\} - \mu_2 \{x \,|x_n \in J\} \leq \Gamma\{(x, y) \,|\, T(x, y) > n\}.$$

Therefore,

$$|\mu_1(X_n \in .) - \mu_2(X_n \in .)|_{tv} \leq 2\, \Gamma\{(x, y) \,|\, T(x, y) > n\}. \tag{4}$$

Proof

$$\mu_1 \{x \,|x_n \in J\} - \mu_2 \{x \,|x_n \in J\} \leq$$

$$\int \int_{x_n \in J} \Gamma(dx, dy) - \int_{y_n \in J} \int \Gamma(dx, dy) =$$

$$[\int_{x_n = y_n} \int_{x_n \in J} \Gamma(dx, dy) + \int_{x_n \neq y_n} \int_{x_n \in J} \Gamma(dx, dy)\,] -$$

$$[\int_{y_n \in J} \int_{x_n = y_n} \Gamma(dx, dy) + \int_{y_n \in J} \int_{x_n \neq y_n} \Gamma(dx, dy)\,] =$$

$$\int_{x_n \neq y_n} \int_{x_n \in J} \Gamma(dx, dy) - \int_{y_n \in J} \int_{x_n \neq y_n} \Gamma(dx, dy) \leq$$

$$\int_{x_n \neq y_n} \int_{x_n \in J} \Gamma(dx, dy) \leq \Gamma\{(x, y) \,|\, T(x, y) > n\},$$

because if x and y are such that $x_n \neq y_n$, then, $T(x, y) > n$. □

We point out that for many plans Γ we have that $T = \infty$ almost everywhere. For some specials ones this is not true.

A more complex result is:

Proposition 4 *Estimation in the Bernoulli space - Suppose* μ_1 *and* μ_2 *are probabilities on* $\{1, 2, .., d\}^{\mathbb{N}}$. *Given* $\Gamma \in \mathscr{C}(\mu_1, \mu_2)$, *then for any n,*

$$|\, (\sigma^n)^*(\mu_1) - (\sigma^n)^*(\mu_2)\,|_{tv} \;\leq 2\, \Gamma\{(x, y) \,|\, T(x, y) > n\}.$$

Proof Remember that if (x, y) is such that for fixed n we have $x_n \neq y_n$, then, $T_\Gamma(x, y) > n$.

Given a set $A \subset \Omega$ in $\mathscr{B}$,

$$(\sigma^n)^*(\mu_1)(A) - (\sigma^n)^*(\mu_2)(A) =$$

$$\mu_1\, \{x\, |(\sigma^n)(x) \in A\} - \mu_2\, \{y\, |(\sigma^n)(y) \in A\} \leq$$

$$\int\int_{\{x\,|\,(\sigma^n)(x)\in A\}} \Gamma(dx, dy) - \int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int \Gamma(dx, dy) =$$

$$[\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}} \Gamma(dx, dy) +$$

$$\int_{\{(\sigma^n)(x)=(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}} \Gamma(dx, dy)\,] -$$

$$[\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}} \Gamma(dx, dy) +$$

$$\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{(\sigma^n)(x)=(\sigma^n)(y)} \Gamma(dx, dy)\,] =$$

$$\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}} \Gamma(dx, dy) -$$

$$\int_{\{y\,|\,(\sigma^n)(y)\in A\}}\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}} \Gamma(dx, dy) \leq$$

$$\int_{\{(\sigma^n)(x)\neq(\sigma^n)(y)\}}\int_{\{x\,|\,(\sigma^n)(x)\in A\}} \Gamma(dx, dy) \leq$$

$$\int\int_{\{(x,y)\,|\,T(x,y)>n\}} \Gamma(dx, dy) \leq \Gamma\{(x, y) \,|\, T(x, y) > n\}.$$

Taking supremum among all A we get the claim. □

Suppose X_n, $n \in \mathbb{N}$ is a stochastic process over S and $\Omega = S^{\mathbb{N}}$. We assume that $X_n : \Omega \to S$ is such that $X_n(w) = w_n$, for any $n \in \mathbb{N}$, where $w = (w_1, w_2, ..., w_n, ..) \in \Omega$. On Ω we consider the sigma-algebra $\mathscr{A}$ generated by the cylinder sets (which is the same as the one generated by the open sets). The stochastic process determines a probability P on the sigma-algebra $\mathscr{A}$ of Ω (see [25] or [17]).

A **stopping time** on Ω is a measurable function $T : \Omega \to \mathbb{N}$, such that, the set $A = \{w \,|\, T(w) = N\}$ depends only $X_1, X_2, , .., X_N$. In other words to know if $T(w) = N$ we just have to consider the string $w_1, w_2, ..., w_N$.

We follow Hollander [5].

Given a plan π on $\Omega \times \Omega$, a **generic stopping time** T and a non-decreasing function $\psi : \mathbb{N} \to [0, \infty)$, such that $\lim_{n\to\infty} \psi(n) = \infty$, assume that

$$\int \psi\,(\,T(x, y)\,)\,\pi(dx, dy) < \infty.$$

Note that for any n

$$\psi(n)\,\pi(T \;>\; n) \;\le \int_{T>n} \psi\,(\,T(x, y)\,)\,\pi(dx, dy).$$

The right hand side tends to zero when $n \to \infty$ by the dominated convergence theorem because $\int \psi\,(\,T(x, y)\,)\,\pi(dx, dy) < \infty$.

Given ε suppose that N is big enough such that for all $n > N$ we have $\psi(n)\,\pi(T \;> n) \le \varepsilon$.

Suppose the plan π is in $\mathscr{C}(\mu_1, \mu_2)$.

Then, from last proposition we have that for n,

$$|\,(\sigma^n)^*(\mu_1) - (\sigma^n)^*(\mu_2)\,|_{tv} \;\le 2\,\pi\{(x, y)\,|\,T(x, y) > n\} \le 2\,\varepsilon\frac{1}{\psi(n)}. \tag{5}$$

Estimates of the form

$$|\mu_1(X_n \in\, .) - \mu_2(X_n \in\, .)|_{tv} \le 2\,\pi\{(x, y)\,|\,T(x, y) > n\} \le 2\,\varepsilon\frac{1}{\psi(n)}. \tag{6}$$

are also important and interesting.

In this way one can get an estimation of the speed of convergence of the above difference by means of the coupling time and ψ. This depends on the plan π we pick. The main point is the smart guess for choosing the plan. All of the above also depends on $\psi(n)$ which can be of polynomial type $n^{-\gamma}$, $\gamma > 0$, or exponential type $e^{-\lambda n}$, $\lambda > 0$, depending of the problem.

The main problem in these kind of questions is to estimate $\pi(T > n)$, where $n \in \mathbb{N}$. This of course depends on π and T.

5 Exponential Convergence to Equilibrium for Finite State Markov Chains

5.1 *The Estimation Using T^1*

We want to consider now the following one: suppose $\mathscr{P}$ is a finite stochastic matrix with all entries positive and λ its unique invariant vector of probability.

Consider a Markov chain X_n with another initial condition ν and the same transition matrix $\mathscr{P}$. We denote the associated probability by P.

We want to show the existence of $0 < \rho < 1$, such that,

$$|\lambda - P(X_n \in .)|_{tv} \leq 2\,(1-\rho)^n.$$

Theorem 2 will describe the speed of convergence (which is exponential) to the equilibrium λ when times n goes to infinity for the chain X_n, $n \in \mathbb{N}$. This result is the main goal of the next subsections.

Definition 8 The coupling time T_1 is **the** measurable function $T_1 : \Omega \times \Omega \to \mathbb{N}$ given by

$$T_1(x, y) = \inf\{n \,|\, x_n = y_n\ \}$$

for any $x, y \in \Omega$.

This value can be eventually ∞.

Note that this coupling time is different from the other one denoted by T. Note also that $T_1(x, y) \leq T(x, y)$ for any (x, y).

We point out that given a plan π on $\Omega \times \Omega$ we have that

$$\pi(T_1(x, y) > n) \leq \pi\{(x, y) \in \Omega \times \Omega \,|\, x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\}.$$

Consider a d dimensional Stochastic Matrix $\mathscr{P} = (P_{i,j})_{i,j=1,2...,d}$ with all entries positive and denote by ρ the minimum value of $P_{i,j}$. Consider two vector of initial probabilities λ and ν. Using the fixed matrix $\mathscr{P}$ they define respectively two different probabilities on Ω which are denoted by μ_1 and μ_2. They generate respectively two different stochastic processes $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ taking values on $\{1, 2..., d\}$. We assume the initial time is $n = 1$. This is consistent with the notation $x = (x_1, x_2, x_3, ...)$.

Consider the plan Γ on $\Omega \times \Omega$ such that $\Gamma = \mu_1 \otimes \mu_2$. For a fixed n we will estimate $\Gamma\{(x, y) \in \Omega \times \Omega \,|\, x_n \neq y_n\}$.

Note that for any others initial vector of probabilities $\tilde{\lambda}$ and $\tilde{\nu}$ we have

$$\Gamma\{(x, y) \in \Omega \times \Omega \,|\, x_n = y_n\} = \sum_{a_n=1}^{d} \Gamma\{(x, y) \in \Omega \times \Omega \,|\, x_n = y_n = a_n\} =$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1}\ P_{a_1,a_2}...P_{a_{n-1},a_n}] [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\nu_j} P_{j,a_1}\ P_{a_1,a_2}..P_{a_{n-1},a_n}] \geq$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1}\ P_{a_1,a_2}...P_{a_{n-1},a_n}] [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\nu_j} P_{j,a_1}\ P_{a_1,a_2}..P_{a_{n-2},a_{n-1}}\ \rho] =$$

$$\sum_{a_n=1}^{d} [\sum_{j,a_1,..a_{n-1}=1}^{d} \tilde{\lambda}_j P_{j,a_1}\ P_{a_1,a_2}...P_{a_{n-1},a_n}]\ \rho = \rho.$$

In this way $\Gamma\{(x, y) \in \Omega \times \Omega \mid x_n \neq y_n\} \leq (1 - \rho)$. A very important remark is that the above expression does not depend of the initial vector of probability $\tilde{\lambda}$ and $\tilde{\nu}$. Indeed, just depend on the matrix $\mathscr{P}$.

As $\Gamma = \mu_1 \otimes \mu_2$, then the associated Stochastic Process $(X_n, Y_n)_{n\in\mathbb{N}}$ taking values on $\{1, 2, ..., d\} \times \{1, 2, ..., d\}$ such that

$$P((X_1, Y_1) \in (A_1, B_1), ..., (X_n, Y_n) \in (A_n, B_n)) =$$

$$\mu_1(A_1 \times ... \times A_n \times \{1, 2, .., d\}^{\mathbb{N}})\ \mu_2(B_1 \times ... \times B_n \times \{1, 2, .., d\}^{\mathbb{N}}) =$$

$$\Gamma(\ (A_1 \times ... \times A_n \times \{1, 2, .., d\}^{\mathbb{N}})\ \times (B_1 \times ... \times B_n \times \{1, 2, .., d\}^{\mathbb{N}})$$

is a Markov chain with stochastic matrix

$$\begin{pmatrix} \mathscr{P} & 0 \\ 0 & \mathscr{P} \end{pmatrix}. \tag{7}$$

The initial vector of probability is such that

$$P(X_1 = i, Y_1 = j)\ =\ \lambda_i\, \nu_j. \tag{8}$$

Γ is Markov probability on the space $\{1, 2, .., d\} \times \{1, 2, .., d\}^{\mathbb{N}}$.

We point out that given any vector of initial probability (do not have to be as above in (8)) for this Markov Chain with transition matrix (7) and with values on $\{1, 2, .., d\} \times \{1, 2, .., d\}$ we get that the associated Markov probability P on the space $(\{1, 2, .., d\} \times \{1, 2, .., d\})^{\mathbb{N}}$ is such that $P(\{(x, y) \in \Omega \times \Omega \mid x_n = y_n\}\,) \geq \rho$.

Note that the event $X_1 \neq Y_1$ is not independent of $X_2 \neq Y_2$.

We want to show that for any n we have

$$\Gamma\{(x, y) \in \Omega \times \Omega \mid x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} \leq (1 - \rho)^n. \tag{9}$$

From this will follow at once that:

Proposition 5

$$\Gamma(T_1 > n) \;\leq\; (1-\rho)^n. \tag{10}$$

Proof Note that

$$\Gamma\{(x,y)\in\Omega\times\Omega\,|\,x_1\neq y_1, x_2\neq y_2, ..., x_n\neq y_n\} =$$

$$\Gamma\{\,X_1\neq Y_1, X_2\neq Y_2, ..., X_n\neq Y_n\} =$$

$$\frac{\Gamma\{\,X_1\neq Y_1, X_2\neq Y_2, ..., X_n\neq Y_n\}}{\Gamma\{X_1\neq Y_1\}}\,\Gamma\{X_1\neq Y_1\} =$$

$$\Gamma\{\,X_1\neq Y_1, X_2\neq Y_2, ..., X_n\neq Y_n\,|\,X_1\neq Y_1\}\,\Gamma\{X_1\neq Y_1\}\leq$$

$$\Gamma\{X_2\neq Y_2, ..., X_n\neq Y_n\,|\,X_1\neq Y_1\}\,(1-\rho).$$

We will need in this moment the following property: suppose Z_n, $n\in\mathbb{N}$, is a Markov chain taking values in a finite set E with transition matrix $\hat{\mathscr{P}}$. Consider also a certain initial condition $\tilde{\pi}$. This defines a Markov Probability P of the space of paths $E^{\mathbb{N}}$.

Then,

$$P(Z_2\in A_2, Z_3\in A_3, ...Z_n\in A_n\,|\,Z_1\in A_1) \;=$$

$$\frac{P(Z_1\in A_1, Z_2\in A_2, Z_3\in A_3, ..., Z_n\in A_n)}{P(Z_1\in A_1)} =$$

$$\frac{\sum_{j\in A_1}\;\tilde{\pi}_{a_j}\,[\sum_{a_2\in A_2}\;...\;\sum_{a_n\in A_n}\;\hat{\mathscr{P}}_{a_j,a_2}\,\hat{\mathscr{P}}_{a_2,a_3}\,...\,\hat{\mathscr{P}}_{a_{n-1},a_n}]}{\sum_{j\in A_1}\tilde{\pi}_{a_j}} =$$

$$P_\gamma(Z_2\in A_2, Z_3\in A_3, ...Z_n\in A_n), \tag{11}$$

where γ is an initial vector of probability on E, such that, for $r\in A_1$ we have

$$\gamma_r = \frac{\tilde{\pi}_{a_r}}{\sum_{j\in A_1}\tilde{\pi}_{a_j}}, \tag{12}$$

and for r which is not in A_1 we have and $\gamma_r = 0$.

The above property is a particular case of Prop 1.7 p. 78 in [19] for a Markov Processes Z_n taking values on E which says

$$P_q(Z_{t+s}\in A\,|\,\mathscr{F}_t) = P_{Z_s}(Z_t\in A),$$

where $\mathscr{F}_t$ is the sigma algebra determined by the process from time 1 to t and q is an initial vector of probability on E. This expression is sometimes called the Markov Property.

We will use (11) taking $Z_n = (X_n, Y_n)$, $A_n = \{X_n \neq Y_n\}$ and $E = \{1, 2, .., d\} \times \{1, 2, .., d\}$.

Therefore,

$$\Gamma\{(x, y) \in \Omega \times \Omega \mid x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} =$$

$$\Gamma\{X_2 \neq Y_2, ..., X_n \neq Y_n \mid X_1 \neq Y_1\}\,(1 - \rho).$$

$$\tilde{\Gamma}\{X_2 \neq Y_2, ..., X_n \neq Y_n\}\,(1 - \rho),$$

where $\tilde{\Gamma}$ is another plan (another Markov probability on the space $\{1, 2, .., d\} \times \{1, 2, .., d\}^{\mathbb{N}}$) due to the fact that we changed the initial condition as described in (11).

Now we use the fact that $\tilde{\Gamma}\{(x, y) \in \Omega \times \Omega \mid x_n \neq y_n\} \leq (1 - \rho)$ and we get that

$$\Gamma\{(x, y) \in \Omega \times \Omega \mid x_1 \neq y_1, x_2 \neq y_2, ..., x_n \neq y_n\} \leq$$

$$\tilde{\Gamma}\{X_2 \neq Y_2, ..., X_n \neq Y_n\}\,(1 - \rho) \leq$$

$$\tilde{\Gamma}\{X_3 \neq Y_3, ..., X_n \neq Y_n \mid X_2 \neq Y_2\}(1 - \rho)^2.$$

Proceeding by induction we get (9). □

Consider a d by d stochastic matrix $\mathscr{P}$ with all positive entries and we denote by X_n, $n \in \mathbb{N}$, and Y_n, $n \in \mathbb{N}$ two independent Markov Processes with values on $\{1, 2..., d\}$ respectively associated to the same matrix $\mathscr{P}$. We denote $S = \{1, 2..., d\}$.

Consider an initial probability λ for X_n^λ and ν for Y_n^ν. This will define probabilities on $\{1, 2..., d\}^{\mathbb{N}}$ which we will denote respectively by P_1 and P_2. We will denote P_1^j the probability we get from the Markov Process defined by the transition matrix $\mathscr{P}$ and the initial vector of probability δ_j, where $j \in \{1, 2, ..., d\}$.

We consider first a Stochastic Process $V_n, n \in \mathbb{N}$, with values on $S^2 = \{1, 2..., d\}^2$ of the form $V_n = (X_n^\lambda, Y_n^\nu)$. This by assumption is the probability $\tilde{P} = P_1 \otimes P_2$ on $(\{1, 2..., d\}^2)^{\mathbb{N}}$. That is the processes X_n^λ and Y_n^ν are independent. By abuse of language $\tilde{P}$ defines a probability on $(\{1, 2..., d\}^2)^n$ for each $n \in \mathbb{N}$.

We consider another Stochastic Process $U_n, n \in \mathbb{N}$, with values on $\{1, 2..., d\}^2$ of the form $U_n = (X_n, Y_n)$. This will define another probability $\hat{P}$ on $(\{1, 2..., d\}^2)^{\mathbb{N}}$. In this case the processes X_n and Y_n are not independent. By abuse of language $\hat{P}$ defines a probability on $(\{1, 2..., d\}^2)^n$ for each $n \in \mathbb{N}$.

For a fixed n we have to define $\hat{P}$ is sets of the form

$$\hat{P}[\,(\,A_1 \times B_1\,) \times (\,A_2 \times B_2\,) \times ... \times (\,A_n \times B_n\,)\, \times (S \times S)^{\mathbb{N}}],$$

where $A_k, B_k \subset S, k = 1, 2..., n$.

We will define $\hat{P}$ in the following way: suppose $T_1 : \{1, 2..., d\}^{\mathbb{N}} \times \{1, 2..., d\}^{\mathbb{N}} \to \mathbb{N}$ is given by

$$T_1(x, y) = \inf\{n \,|\, x_n = y_n \}$$

for any $x, y \in \{1, 2..., d\}^{\mathbb{N}}$.

Consider a fixed value n.

Denote by G_n the set of elements on $\{1, 2..., d\}^n \times \{1, 2..., d\}^n$ such that $T_1(x, y) > n$.

For any subset $K \subset G_n$ define

$$\hat{P}(K \times (S \times S)^{\mathbb{N}}) = \tilde{P}(K \times (S \times S)^{\mathbb{N}}). \tag{13}$$

This defines $\hat{P}$ on cylinders of the form

$$(\{a_1\} \times \{b_1\}) \times (\{a_2\} \times \{b_2\}) \times ... \times (\{a_n\} \times \{b_n\}) \times (S \times S)^{\mathbb{N}} \subset G_n \times (S \times S)^{\mathbb{N}},$$

$n \in \mathbb{N}$, $a_j, b_j \in \{1, 2..., d\}$, $j = 1, 2..., n$.

For a fixed n the sets of the form $H_k^n = H_k = \{T_1 = k\}, k = 1, 2, ...n$, and the set $G_n \times (S \times S)^{\mathbb{N}}$ define a partition of $\{1, 2..., d\}^{\mathbb{N}} \times \{1, 2..., d\}^{\mathbb{N}}$.

We denote by Δ the diagonal on $\{1, 2..., d\}^2$.

For a fixed n and $k \leq n$, now we will define $\hat{P}$ on subsets of H_k^n.

This means we suppose that k is the smaller one among $\{1, 2, ..., n\}$, such that, $(A_k \times B_k) \cap \Delta \neq \emptyset$.

For defining probabilities on this case we can assume without lost of generality that $A_k = B_k = \{j\}$ for a certain j.

Then,

$$\hat{P}[(A_1 \times B_1) \times (A_2 \times B_2) \times ... \times (A_n \times B_n)] =$$

$$P_1(A_1 \times ... \times A_{k-1} \times j)\; P_2(B_1 \times ... \times B_{k-1} \times j)\; P_1^j(A_{k+1} \times A_{k+2} \times ... \times A_n). \tag{14}$$

This procedure defines a probability $\hat{P}$ on $(\{1, 2, ..., d\}^2)^n$ and subsequently on $(\{1, 2, ..., d\}^2)^{\mathbb{N}}$.

One can say in an elusive way that $\hat{P}$ describes the following process: two particles located on S evolve in time in an independent way (a Markov Chain in $S \times S$) and when they meet they stay together in the future according to the law of the Markov chain in S.

Note that for fixed n and $j \in S$, we have

$$\hat{P}\{(X_n, Y_n) = (j, j), \text{ and also } (X_m, Y_m) \cap \Delta = \emptyset\,, \text{ for some } n < m \} = 0. \tag{15}$$

We claim that:

Proposition 6 *For fixed $j \in \{1, 2, .., d\}$ and $n \in \mathbb{N}$*

$$\hat{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = P_1(X_n = j). \tag{16}$$

Proof We denote $S = \{1, 2, ..., d\}$ and we consider a fixed $n \in \mathbb{N}$ and a fixed $j \in S$.

Note that

$$\tilde{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = \sum_{r=1}^{d} \tilde{P}(X_n = j, Y_n = r) =$$

$$P_1(X_n = j) \sum_{r=1}^{d} P_2(Y_n = r) = P_1(X_n = j). \tag{17}$$

Given $(x, y) \in (S \times S)^{\mathbb{N}}$ we have that either $T_1(x, y) > n$ or $T_1(x, y) \leq n$.

For a fixed n we use the notation described above for the sets of the form $H_k^n = H_k = \{T_1 = k\}$, $k = 1, 2, ...n$, and the set $G_n \times (S \times S)^{\mathbb{N}}$ which define a partition of $\{1, 2..., d\}^{\mathbb{N}} \times \{1, 2..., d\}^{\mathbb{N}}$.

Note also that

$$\hat{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = \sum_{r \neq j}^{d} \hat{P}(X_n = j, Y_n = r) + \hat{P}(X_n = j, Y_n = j) =$$

$$\sum_{r \neq j}^{d} \tilde{P}(X_n = j, Y_n = r) + \hat{P}(X_n = j, Y_n = j). \tag{18}$$

From (15) and also (13) we get that $\sum_{r \neq j}^{d} \tilde{P}(X_n = j, Y_n = r) = \sum_{r \neq j}^{d} \hat{P}(X_n = j, Y_n = r)$. This part corresponds to the subset $G_n \times (S \times S)^{\mathbb{N}}$.

We claim that $\hat{P}(X_n = j, Y_n = j) = \tilde{P}(X_n = j, Y_n = j)$.

Indeed,

$$\hat{P}(X_n = j, Y_n = j) =$$

$$\sum_{k=1}^{n-1} \hat{P}(\{T_1 = k\} \cap (S \times S)^n) + \hat{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{r=1}^{d} \sum_{k=1}^{n-1} \hat{P}((S \times S)^{k-1} \times (\{r\} \times \{r\}) \times (S \times S)^{n-k}) +$$

$$\hat{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{r=1}^{d} \sum_{k=1}^{n-1} \hat{P}((S \times S - \Delta)^{k-1} \times (\{r\} \times \{r\}) \times (S \times S)^{n-k}) +$$

$$\hat{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{r=1}^{d} \sum_{k=1}^{n-1} \tilde{P}((S \times S - \ \Delta)^{k-1} \times (\{r\} \times \{r\}) \times (S \times S)^{n-k}) +$$

$$\tilde{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\sum_{k=1}^{n-1} \tilde{P}(\{T_1 = k\} \cap (S \times S)^n) + \tilde{P}(\{T_1 = n\} \cap \{(X_n, Y_n) = (j, j)\}) =$$

$$\tilde{P}(X_n = j, Y_n = j).$$

Now, from (18) we get that

$$\hat{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = \tilde{P}(X_n = j, Y_n \in \{1, 2, .., d\}) = P_1(X_n = j).$$

□

In a similar way we get:

Proposition 7 *For fixed $j \in \{1, 2, .., d\}$ and $n \in \mathbb{N}$*

$$\hat{P}(X_n \in \{1, 2, .., d\}, Y_n = j,) = P_2(Y_n = j). \tag{19}$$

We will need later the following result:

Proposition 8 *The probability $\hat{P}$ on $(\{1, 2..., d\}^2)^{\mathbb{N}}$ which is a plan on the product space $\{1, 2..., d\}^{\mathbb{N}} \times \{1, 2..., d\}^{\mathbb{N}}$ satisfies*

$$\hat{P}(T > n) = \hat{P}(T_1 > n) = \tilde{P}(T_1 > n), \tag{20}$$

where $\tilde{P}$ is the independent process.

Proof We have that $\hat{P}(T > n) = \hat{P}(T_1 > n)$ because of (15).

Finally, we claim that $\hat{P}(T_1 > n) = \tilde{P}(T_1 > n)$. Indeed, if $m > n$ we have from (13) that $\hat{P}(\{T_1 = m\} \cap (S \times S)^m) = \tilde{P}(\{T_1 = m\} \cap (S \times S)^m)$. □

5.2 Speed Estimation on Total Variation

The probabilities P_1 and P_2 on $\{1, 2, ..., d\}^{\mathbb{N}}$ and X_n, Y_n, $n \in \mathbb{N}$, were described above. We want to take advantage of (4), (10) and (16).

We get before estimates of the form (6) for T satisfying expression (4), that is, for X_n and Y_n independent

$$|P_1(X_n \in \ .) - P_2(Y_n \in .)|_{tv} \leq 2\, \Gamma\{(x, y) \,|\, T(x, y) > n\},$$

when

$$T(x, y) = \inf\{n \,|\, x_m = y_m \text{ for all}\, m \geq n\},$$

and where Γ satisfies the Hypothesis (2) and (3) of Proposition 3.

This of course depends of a good choice of Γ. We take $\Gamma = \hat{P}$ and we know from (16) and (19) that these hypothesis are true.

Then, we get from (20) that

$$|P_1(X_n \in \ .) - P_2(Y_n \in .)|_{tv} \leq 2\, \hat{P}\{(x, y) \,|\, T > n\} = 2\, \tilde{P}\{(x, y) \,|\, T_1 > n\}.$$

Now, from (10) we get that $\tilde{P}(T_1 > n) \ \leq \ (1 - \rho)^n$.

From this follows that

$$|P_1(X_n \in \ .) - P_2(Y_n \in .)|_{tv} \leq \ 2\,(1 - \rho)^n. \tag{21}$$

Note that $P_1(X_n \in \ .)$ and $P_2(Y_n \in \ .)$ are probabilities on the state space S.

The bottom line is: suppose λ is the stationary vector for the d by d matrix $\mathscr{P}$ (which have all entries positive) and ν is another initial vector of probability on $\{1, 2, ..., d\}$. This defines respectively two probabilities on $\{1, 2, ..., d\}^{\mathbb{N}}$ which we denote P_1 and P_2. We also denote, respectively, $X_n, n \in \mathbb{N}$, the first Markov Process and $Y_n, n \in \mathbb{N}$, the second.

Note that for any n and $J \subset S$ we have $P_1(X_n \in J) = P_1(X_0 \in J)$ because λ is stationary. In other words $P_1(X_n \in \ .)$ is λ

In this way we are analyzing the time evolution of two probabilities on the space state S and from (21), we get:

Theorem 2 *Suppose the transition matrix $\mathscr{P}$ has all entries positive and $\lambda\, \mathscr{P} = \lambda$, where λ is the initial stationary vector of probability on S. Then, for any n we have*

$$|\lambda - P_2(Y_n \in .)|_{tv} \leq \ 2\,(1 - \rho)^n,$$

and this describes the speed of convergence to the equilibrium λ when times goes to infinity for a chain Y_n with initial condition ν and matrix transition $\mathscr{P}$.

6 The $\bar{d}$ distance

Given two probabilities on μ, ν in the set $\Omega = \{1, 2\}^{\mathbb{N}}$, consider the set $\mathscr{C}(\mu, \nu)$ of plans λ in $\Omega \times \Omega$ such the projection in the first coordinate is μ and in the second is ν.

Definition 9 A joining is a probability on $\mathscr{C}(\mu, \nu)$ which is invariant for the dynamical system $T : \Omega \times \Omega \to \Omega \times \Omega$ given by $T(x, y) = (\sigma(x), \sigma(y))$. The set of such joinings is denoted by $J(\mu, \nu)$.

A general reference for joinings is [11].

Definition 10 Denote $\{1, 2\}^n = Q_n$ and consider the d_n-Hamming metric in Q_n defined by

$$d_n(x, y) = \frac{1}{n} \sum_{j=1}^{n} I_{\{x_j \neq y_j\}}.$$

Points x in Ω are denoted by $x = (x_0, x_1, x_2, ...)$.

Definition 11 For two σ-invariant probabilities μ and ν we define the distance

$$\bar{d}(\mu, \nu) = \inf_{\lambda \in J(\mu,\nu)} \int I_{\{x_0 \neq y_0\}} \lambda(dx, dy) =$$

$$\inf_{\lambda \in J(\mu,\nu)} \lim_{n \to \infty} \int d_n(x, y) \lambda(dx, dy) =$$

$$\lim_{n \to \infty} \inf_{\lambda \in J(\mu,\nu)} \int d_n(x, y) \lambda(dx, dy).$$

In the above equalities we used the ergodic theorem (see [25]).

The above value try to minimize the asymptotic mean disagreement between the symbols of the paths for a given plan.

We point out that the $\bar{d}$ has a dynamical content. Several dynamical properties are preserves under limit over the $\bar{d}$ distance (see [4]). In this paper properties related to Bernoullicity, g-measures and Gibbs states are considered. For instance the map that takes a potential to its equilibrium state is continuous with respect to the $\bar{d}$ distance (see Theorem 3 in [4]).

Theorem 3 *Suppose μ is the independent Bernoulli probability associated to p_1, p_2 and ν is the independent Bernoulli probability associated to q_1, q_2.*

Suppose $p_1 \leq q_1$. Then,

$$\bar{d}(\mu, \nu) = q_1 - p_1 = \frac{1}{2}(|q_1 - p_1| + |q_2 - p_2|).$$

Proof Given $\lambda \in J(\mu, \nu)$ we have

$$\int I_{\{x_0 \neq y_0\}} \lambda(dx, dy) =$$

$$\int_{[1]\times[1]} I_{\{x_0 \neq y_0\}} \lambda(dx, dy) + \int_{[1]\times[2]} I_{\{x_0 \neq y_0\}} \lambda(dx, dy) +$$

$$\int_{[2]\times[1]} I_{\{x_0\neq y_0\}}\lambda(dx,dy) + \int_{[2]\times[2]} I_{\{x_0\neq y_0\}}\lambda(dx,dy) =$$

$$\int_{[1]\times[2]} I_{\{x_0\neq y_0\}}\lambda(dx,dy) + \int_{[2]\times[1]} I_{\{x_0\neq y_0\}}\lambda(dx,dy).$$

We denote $a_{i,j} = \int_{[i]\times[j]} I_{\{x_0\neq y_0\}}\lambda(dx,dy)$.

We have the relations

$$\lambda_{1,1} + \lambda_{2,1} = \nu[1] = q_1\,,\ \ \lambda_{1,2} + \lambda_{2,2} = \nu[2] = q_2\,,$$

$$\lambda_{1,1} + \lambda_{1,2} = \mu[1] = p_1\,,\ \ \lambda_{2,1} + \lambda_{2,2} = \mu[2] = p_2\,.$$

Denote by $t = \lambda_{1,1}$, then

$$\lambda_{1,2} = p_1 - t,$$

$$\lambda_{2,1} = q_1 - t,$$

and

$$\lambda_{2,2} = p_2 - q_1 + t,$$

We have to minimize $a_{12} + a_{21} = p_1 + q_1 - 2t$ given the above linear constrains. This is a linear maximization problem and the largest possible value of t will be the solution.

From the fact that the $\lambda_{i,j}$ are non negative we get that $t \leq q_1$ and $t \leq p_1$. But as we assume that $q_1 \geq p_1$, we get the only restriction $0 \leq t \leq p_1$.

Taking $\lambda_{11} = p_1$ we get the minimal value for the sum $a_{12} + a_{21}$ which is $p_1 + q_1 - 2p_1 = q_1 - p_1$.

This shows that $\bar{d}(\mu,\nu) \geq q_1 - p_1$.

Now we will show that there exists a plan λ that realizes the value $q_1 - p_1$.

Consider a new Bernoulli independent process (X_n, Y_n) with value on $\{1,2\} \times \{1,2\}$.

We set

$$P(X_0 = 1, Y_0 = 1) = p_1,\ P(X_0 = 1, Y_0 = 2) = \ 0,$$

$$P(X_0 = 2, Y_0 = 1) = q_1 - p_1,\ P(X_0 = 2, Y_0 = 2) = q_2 = p_2 - q_1 + p_1,$$

It is easy to see that such plan λ is in $J(\mu,\nu)$ and $\int I_{\{x_0\neq y_0\}}\lambda(dx,dy) = q_1 - p_1$. □

A natural question is what can be said for the $\bar{d}$ distance of two finite state Markov Processes (taking values in the same state space S) obtained from two different stochastic matrices, or more generally, for Gibbs states. This is a not so easy problem and partial results can be found for instance in [2, 3, 6, 10]. In most of them couplings are used in an essential way.

We refer the reader to [8, 15, 22] for more details on couplings from a probabilistic point of view.

7 Contraction for the Dual of the Ruelle Operator

In this section we study properties related to Gibbs states of Lipchitz potentials (see [18] for general results).

Suppose $A = \log J$ is Lipchitz and normalized, that is, for any $x \in \Omega$ we have $\mathscr{L}_{\log J}(1)(x) = 1$.

If $x = (x_1, x_2, ..) \in \{1, 2, ..., d\}^{\mathbb{N}}$ and $t \in \mathbb{N}$, we denote by $z_j^t(x)$, $j = 1, 2, ..., d^t$, the d^t solutions of $\sigma^t(z) = x$.

We denote $J^t(z_j^t(x))$ the expression $e^{\sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(x))}$.

Therefore, given x and y we have two probabilites

$$(P^t)^*(\delta_x) = \sum_{j=1}^{d^t} \delta_{z_j^t(x)} J^t(z_j^t(x))$$

and

$$(P^t)^*(\delta_y) = \sum_{j=1}^{d^t} \delta_{z_j^t(y)} J^t(z_j^t(y)).$$

For each k we have that the points $z_j^k(x)$ and $z_j^k(y)$, $j = 1, 2, 3, .., 2^k$, are all different but the distance $d_\theta(z_j^k(x), z_j^k(y))$ is at most θ^k (j pair by pair).

Suppose $A = \log J$ has Lipchitz constant M. Then, for t and $j = 1, 2, 3, .., 2^t$ fixed

$$| \sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(x)) - \sum_{k=0}^{t-1} \log J(\sigma^k(z_j^t(y)) | \leq$$

$$\sum_{k=0}^{t-1} M\, d_\theta(\sigma^k(z_j^t(x)) - \sigma^k(z_j^t(y))\,)\,| \leq$$

$$M\,[\sum_{k=0}^{t-1} \theta^k\,]\, d_\theta(x, y) \leq M \frac{1}{1-\theta}\, d_\theta(x, y)$$

In this way for any x and y we have for any t, $j = 1, 2, 3, .., 2^t$,

$$\frac{J^t(z_j^t(x))}{J^t(z_j^t(y))} \le e^{M\,\frac{1}{1-\theta}\,d_\theta(x,y)}. \tag{22}$$

The above kind of estimation is known as the bounded distortion property. It is key element in the subsequent developments.

Lemma 1 *For every $\delta > 0$, there is T such that for any $k \ge T$ there exists an $a > 0$, so that*

$$\sup_{\Gamma\in\mathscr{C}(\,(P^k)^*(\delta_x),\,(P^k)^*(\delta_y))} \Gamma\,\{(x',y'))\in\,\Omega\times\Omega\;:\;d_\theta(x',y')\le\delta\,\}\ge a.$$

Proof In order to proof the Lemma, given x and y we construct explicitly an element in

$$\mathscr{C}(\,(P^k)^*(\delta_x),\;(P^k)^*(\delta_y)).$$

By the orbit structure of the Bernoulli shift, the preimages of x and y come in pairs, and as stated above, the distance of a pair satisfies $d_\theta(z_j^k(x), z_j^k(y)) < \theta^k$. Hence, if $T > \log\delta/\log\theta$, then $d_\theta(z_j^T(x), z_j^T(y)) < \delta$. The other basic observation for the construction stems from bounded distortion in (22). Namely, there exists $\Lambda \in (0,1]$, independent from T, j and x such that[1] for $j = 1, 2, 3, .., d^T$

$$\Lambda J^T(z_j^T(x)) \le \alpha_j^T := \inf_{z\in\Omega} J^T(z_j^T(z)) \le J^T(z_j^T(x)).$$

For $\beta_j^T(x) := J^T(z_j^T(x)) - \alpha_j^T$, $j = 1, 2, 3, .., d^T$, we hence obtained a decomposition into a strictly positive part $\beta_j^T(x)$ and a strictly positive part α_j^T, which is independent from x (and y) but comparable to $J^T(z_j^T(x))$ by Λ. Hence, we obtain a decomposition of a probability measure into two sub-probability measures by

$$(P^T)^*(\delta_x) = \sum_{j=1}^{d^T}\delta_{z_j^T(x)}\alpha_j^T + \sum_{j=1}^{d^T}\delta_{z_j^T(x)}\beta_j^T(x) =: \mu_T + \nu_T^x =$$

$$\sum_{j=1}^{d^T}\delta_{z_j^T(x)}J^T(z_j^T(x)).$$

We claim that the probability Γ on $\Omega\times\Omega$ defined by

$$\Gamma := \sum_{j=1}^{d^T}\delta_{(z_j^T(x),z_j^T(y))}\alpha_j^T + \frac{1}{\nu_T^x(\Omega)}\nu_T^x\otimes\nu_T^y.$$

[1] $\Lambda = 1$ if and only if J is constant on cylinders of length 1.

is in $\Gamma \in \mathscr{C}(\,(P^k)^*(\delta_x),\ (P^k)^*(\delta_y))$.

Indeed, consider a Borel set A then

$$\Gamma(\Omega \times A) = \sum_{j=1}^{d^T} \delta_{(z_j^T(x), z_j^T(y))} \alpha_j^T (\Omega \times A) + \frac{1}{\nu_T^x(\Omega)} \nu_T^x \otimes \nu_T^y\, (\Omega \times A) =$$

$$\sum_{j=1}^{d^T} \delta_{z_j^T(y)} \alpha_j^T (A) + \nu_T^y(A) = \mu_T(A) + \nu_T^y(A) =$$

$$\sum_{j=1}^{d^T} \delta_{z_j^T(y)} J^T(z_j^T(y))\,(A) = (P^T)^*(\delta_y)\,(A).$$

Note that $\nu_T^x(\Omega) = \nu_T^y(\Omega) = 1 - \mu_T(\Omega)$.

In the same way given A we have

$$\Gamma(A \times \Omega) = \sum_{j=1}^{d^T} \delta_{(z_j^T(x), z_j^T(y))} \alpha_j^T (A \times \Omega) + \frac{1}{\nu_T^x(\Omega)} \nu_T^x \otimes \nu_T^y\, (A \times \Omega) =$$

$$\sum_{j=1}^{d^T} \delta_{z_j^T(x)} \alpha_j^T (A) + \nu_T^x(A)\, \frac{\nu_T^y(\Omega)}{\nu_T^x(\Omega)} = \mu_T(A) + \nu_T^x(A) =$$

$$\sum_{j=1}^{d^T} \delta_{z_j^T(x)} J^T(z_j^T(x))\,(A) = (P^T)^*(\delta_x)\,(A).$$

The analogous result for sets of the for $\Omega \times A$ is true.

This shows that $\Gamma \in \mathscr{C}(\,(P^k)^*(\delta_x),\ (P^k)^*(\delta_y))$.

We claim that Γ also satisfies

$$\Gamma\,\{(x', y')) \in\ \Omega \times \Omega\ :\ d_\theta(x', y') \leq \delta\,\} \geq \Lambda.$$

Indeed,

$$\Gamma\,\{(x', y')) \in\ \Omega \times \Omega\ d_\theta(x', y') \leq \delta\,\} \geq$$

$$\sum_{j=1}^{d^T} \delta_{(z_j^T(x), z_j^T(y))}\ \alpha_j^T\ [\ \{(x', y')) \in\ \Omega \times \Omega\ :\ d_\theta(x', y') \leq \delta\,\}\ ] =$$

$$\sum_{j=1}^{d^T} \delta_{z_j^T(x)}\ \alpha_j^T(x) \geq \sum_{j=1}^{d^T} \Lambda J^T(z_j^T(x)) \ =\ \Lambda$$

This proves the Lemma for $a := \Lambda$. □

We define

$$\mathrm{var}_n f\ = \sup\{|f(x) - f(y)|\ : x_i = y_i,\ 0 \leq i\ < n\},$$

and the pseudo norm

$$|f|_\theta = \sup\{\frac{\mathrm{var}_n f}{\theta^n}\ \ : n \geq 0\}.$$

We know that if $\log J$ is Lipchitz, then, there exist $C > 0$ and $\alpha_1 \in (0,1)$ such that for any $t \geq 0$, and any Lipchitz function ϕ we have (prop 2.1 in [18])

$$|\mathscr{L}^t(\phi)\,|_\theta \leq C\ \sup_x |\phi(x)| + \theta^t |\phi|_\theta.$$

We wrote this in the notation of [12] as

$$\sup\{\frac{\mathrm{var}_n\ \mathscr{L}^t(\phi)}{\theta^n}\ \ : n \geq 0\} \leq\ C\ \sup_x |\phi(x)|\ +\ \alpha_1 \sup\{\frac{\mathrm{var}_n\ \phi}{\theta^n}\ \ : n \geq 0\}.$$

The above expression is known as the Lasota–Yorke inequality.

Take $\delta < \frac{1-\alpha_1}{2\,C}$. Now, we define a metric $d(x,y) = \min\{1, \delta^{-1} d_\theta(x,y)\}$. The two metrics d and d_θ are equivalent.

Remember that we denote $\mathscr{C}(\mu_1,\mu_2)$ the set plans in $\Omega \times \Omega$ such that the projection in the first coordinate is μ_1 and in the second is μ_2.

Remember also that, we define the 1-Wasserstein metric associated to d

$$d_1(\mu_1,\mu_2)\ = \inf\{\int\int d(x,y)\ d\ \Gamma(dx,dy)\ |\ \Gamma \in \mathscr{C}(\mu_1,\mu_2)\,\}.$$

Proposition 9 *There exist* $\alpha < 1$*, where* $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$*, and* $t > 0$*, such that, for any* x, y

$$d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y)) \leq \alpha\, d(x,y).$$

The proof will be done later.

Suppose that this result is proved then we get:

Theorem 4 *There exist* $\alpha\ < 1$*, where* $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1+\alpha_1)\}$*, and* $t > 0$*, such that, for any* μ_1, μ_2

$$d_1((P^t)^*(\mu_1),(P^t)^*(\mu_2)) \leq \alpha\, d_1(\mu_1,\mu_2).$$

Proof The main idea is to prove first the following claim: suppose Q is the d_1-optimal plan for μ_1 and μ_2, then,

$$d_1\,((P^t)^*(\mu_1),(P^t)^*(\mu_2)) \leq \int d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y))\, dQ(dx,dy).$$

Suppose the plan in $\Omega \times \Omega$ denoted by $Q(dx,dy)$ has marginals μ_1 and μ_2 in respectively the first and second coordinates.

We will prove the result for a more general continuous potential c. Then, you just have to take $c = d$ in order to get the claim.

Given a continuous cost $c(z_1,z_2), c : X \times X \to \mathbb{R}$, we assume that Q is c-optimal for μ_1 and μ_2. Now, given two points x, y suppose $R^{x,y}(dz_1,dz_2)$ is the c-optimal probability plan for $P^*(\delta_x)$ and $P^*(\delta_y)$.

We denote $S(d\,z_1,d\,z_2)$ the plan

$$S(dz_1,dz_2) = \int\int R^{x,y}(dz_1,dz_2)\; Q(dx,dy).$$

We are going to show that the marginals of this plan are μ_1 and μ_2.

Indeed,

$$\int\int \varphi(z_1)S(dz_1,dz_2) =$$

$$\int\int\int\int \varphi(z_1)R^{x,y}(dz_1,dz_2)\; Q(dx,dy)\,dx\,dy\,dz_1\,dz_2 =$$

$$\int\int\int \varphi(z_1)P^*(\delta_x)\; Q(dx,dy)\,dxdydz_1 =$$

$$\int [\,P^*\,(\int\int \varphi(z_1)\; Q(\,dx\;,dy)dydz_1\,)\,]\,(\delta_x) =$$

$$\int [\,P^*\,(\int \varphi(.)\; Q(\,.\,,dy)dy\,)\,(\delta_x)\,]\,dx = \int \varphi(x)\;\, dP^*(\mu_1)\;\,(dx)$$

because P^* is linear on measures.

In this way the first marginal of $S(dz_1,dz_2)$ is $P^*\,(\mu_1)$.

In the same way one can prove that the second marginal of $S(dz_1,dz_2)$ is $P^*\,\mu_2$. Now we consider $c(x,y) = d(x,y)$.

From the above we get

$$d_1(P^*\,(\mu_1),\,P^*\,(\mu_2)) \le \int d(x,y)\,S(dx,dy),$$

where S was defined from Q which is the d-optimal plan for μ_1 and μ_2.

Therefore, from the above

$$d_1\,(P^*(\mu_1),\,P^*(\mu_2)) \le \int\, d_1(P^*(\delta_x),\,P^*(\delta_y))\,dQ(dx,dy),$$

where Q is the d_1-optimal plan for μ_1 and μ_2.

In a similar way given $t>0$ on can show the analogous result

$$d_1\,((P^t)^*(\mu_1),(P^t)^*(\mu_2)) \le \int\, d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y))\,dQ(dx,dy),$$

where Q is the d_1-optimal plan for μ_1 and μ_2.

Therefore, if there exists a $t>0$ and $\alpha<1$, such that, for any x and y we have

$$d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y)) \,\le\, \alpha\, d(x,y),$$

then

$$d_1((P^t)^*(\mu_1),(P^t)^*(\mu_2) \,\le\, \alpha\ d_1(\mu_1,\mu_2).$$

This is so because

$$d_1\,((P^t)^*(\mu_1),(P^t)^*(\mu_2)) \le \int\, d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y))\,dQ(dx,dy) \le$$

$$\alpha \int\, d(x,y)\,dQ(dx,dy) = \alpha\ d_1(\mu_1,\mu_2).$$

□

Now we prove that for any x,y

$$d_1((P^t)^*(\delta_x),(P^t)^*(\delta_y)) \le \alpha\, d(x,y).$$

Suppose $d_\theta(x,y)\le\delta$, where $\delta<\frac{1-\alpha_1}{2C}$.
Remember that $d(x,y)=\min\{1,\delta^{-1}d_\theta(x,y)\}$. In this case $d(x,y)=\delta^{-1}d_\theta(x,y)$.
By Kantorovich duality (see [23, 24])

$$d_1(\mu_1,\mu_2)\,=\sup_{\phi:X\to\mathbb{R}\text{ has }d\text{ Lipchitz constant }\le 1}\{\int \phi\,d\,\mu_1 - \int \phi\,d\mu_2\,\}.$$

We have to show that: if ϕ has d-Lipchitz constant smaller than 1, then, for such pair of x, y

$$|\mathscr{L}^t_{\log J}\phi(x) - \mathscr{L}^t_{\log J}\phi(y)| \leq \alpha d(x, y) = \alpha\delta^{-1} d_\theta(x, y).$$

We can assume without lost of generality that ϕ attains the value 0.

In this case $\sup_x |\phi(x)| \leq 1$.

Moreover,

$$\sup \frac{\delta\ |\phi(x) - \phi(y)|}{d_\theta(x, y)} \leq \sup \frac{|\phi(x) - \phi(y)|}{d(x, y)} \leq 1.$$

Then,

$$\frac{|\mathscr{L}^t_{\log J}(\phi)(x) - \mathscr{L}^t_{\log J}(\phi)(y)|}{d_\theta(x, y)} \leq$$

$$C \sup_x |\phi(x)| + \alpha_1 \sup \frac{|\phi(x) - \phi(y)|}{d_\theta(x, y)} \leq C + \alpha_1\, \delta^{-1}.$$

As $\delta < \frac{1-\alpha_1}{2C}$, then $C < \frac{1-\alpha_1}{2}\,\delta^{-1}$.

Therefore, from the above, we get that for any x, y such that $d_\theta(x, y) \leq \delta$, we have

$$|\mathscr{L}^t_{\log J}(\phi)(x) - \mathscr{L}^t_{\log J}(\phi)(y)| \leq d_\theta(x, y)\,(C + \alpha_1\,\delta^{-1}) \leq$$

$$d_\theta(x, y)\,(\frac{1-\alpha_1}{2}\,\delta^{-1} + \alpha_1\,\delta^{-1}) =$$

$$d_\theta(x, y)\,\delta^{-1}\,(\frac{1+\alpha_1}{2}) = d(x, y)\ (\frac{1+\alpha_1}{2}) \leq d(x, y)\,\alpha,$$

because $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1 + \alpha_1)\}$.

Now we suppose that x, y such that $d_\theta(x, y) > \delta$. This implies that $d(x, y) = 1$.

We denote

$$\Delta_\delta = \{(x', y') \in\ \Omega \times \Omega\ :\ d_\theta(x', y') \leq \frac{1}{2}\delta\,\}.$$

For such δ there exists $a > 0$ and $T > 0$, such that, for $k > T$, there exists a plan $\Gamma = \Gamma_k$ which satisfies $\Gamma \in \mathscr{C}(\,(P^k)^*(\delta_x),\ (P^k)^*(\delta_y))$ and $\Gamma\,(\Delta_\delta)\ \geq a$.

Note that if $d_\theta(x', y') \leq \frac{1}{2}\delta$, then $d(x', y') \leq \frac{1}{2}$.

Therefore,

$$\int d(x', y')\,\Gamma(dx', dy')\ \ \leq \frac{1}{2}\Gamma\,(\Delta_\delta) + 1 - \Gamma\,(\Delta_\delta) = 1 - \frac{1}{2}\Gamma\,(\Delta_\delta) \leq 1 - \frac{a}{2},$$

because $d(x', y') \leq 1$ in the complement of Δ_δ.

Then, if $d_\theta(x, y) > \delta$ we get

$$d_1((P^t)^*(\delta_x), (P^t)^*(\delta_y)) \leq \int d(x', y')\, \Gamma(dx', dy') \;\leq$$

$$(1 - \frac{a}{2}) = (1 - \frac{a}{2})\; d\,(x, y) \;\leq\; \alpha\; d(x, y)$$

because $\alpha = \max\{1 - \frac{a}{2}, \frac{1}{2}(1 + \alpha_1)\}$.

From all this it follows the main result.

References

1. Austin, T.: Egodic Theory - Notes 11: Entropy, couplings and joinings. Lecture Notes - Courant Institute, NYU (2013)
2. Bressaud, X., Fernandez, R., Galves, A.: Decay of correlations for non-Holderian dynamics. a coupling approach. Electron. J. Probab. **4**(3), 19 (1999). electronic
3. Bressaud, X., Fernandez, R., Galves, A.: Speed of $\bar{d}$-convergence for Markov approximations of chains with complete connections. a coupling approach, Stoch. Process. Appl. **83**, 127–138 (1999)
4. Coelho, Z., Quas, A.: A criteria for $\bar{d}$-continuity. Trans. AMS **350**(8), 3257–3268 (1998)
5. den Hollander, F.: Probability Theory: The Coupling Method. Leiden University, Lectures Notes - Mathematical Institute (2012)
6. Ellis, M.: Distances between two-state Markov Processes attainable by Markov Joinings, TAMS, vol. 241. (1978)
7. Fernandez, R., Ferrari, P., Galvez, A.: Coupling, Renewal and Perfect Simulation of Chains of Infinite Order, University of Rouen (2001)
8. Ferrari, P., Galves, A.: Construction of Stochastic Processes, Coupling and Regeneration, XIII Escuela Venezolana de Matematica
9. Galatolo, S., Pacifico, M.J.: Lorenz-like flows: exponential decay of correlations for the Poincar map, logarithm law, quantitative recurrence, ETDS 30, no. 6, 17031737. (2010)
10. Gallo, S., Lerasle, M., Takahashi, D.: Markov approximation of chains of infinte order in the $\bar{d}$ metric. Markov Process. Relat. Fields **19**(1), 51–82 (2013)
11. Glasner, E.: Ergodic Theory via Joinings. AMS, Providence (2003)
12. Hairer, M., Mattingly, J.: Spectral gaps in Wasserstein distance and the 2-D stochastic Navier–Stokes equations. Ann. Probab. **36**(6), 2050–2091 (2008)
13. Kloeckner, B.: Optimal transport and dynamics of expanding circle maps acting on measures. Ergod. Theory Dyn. Sys. **33**(2), 529–548 (2013)
14. Kloeckner, B., Lopes, A.O., Stadlbauer, M.: Contraction in the Wassertein metric for some Markov Chains and applications for the dynamics of expanding maps. Nonlinearity **28**(11), 4117–4137 (2015)
15. Levin, D., Perez, Y., Wilmer, E.: Markov Chains and Mixing Times. AMS, Providence (2008)
16. Lindvall, T.: Lectures on the Coupling Method. Dover, New York (1992)
17. Lopes, A., Lopes, S.: Introdução aos Processos Estocásticos para estudantes de Matemática, UFRGS (2015)
18. Parry, W., Pollicott, M.: Zeta functions and the periodic orbit structure of hyperbolic dynamics, Astérisque vol. 187–188 (1990)
19. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Springer, Berlin (1991)
20. Stadlbauer, M.: Coupling methods for random topological Markov chains, to appear in Ergodic Theory and Dynamical Systems

21. Sulku, H.: Explicit correlation bounds for expanding maps using the coupling method (2013)
22. Torrison, H.: Coupling, Stationary and Regeneration. Springer, Berlin (2000)
23. Villani, C.: Topics in Optimal Transportation. AMS, Providence (2003)
24. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2009)
25. Walkden, C.: Ergodic Theory, Lecture Notes University of Manchester (2014)

Extreme Weather, Biotechnology, and Corn Productivity

Jonathan R. McFadden and John A. Miranowski

Abstract U.S. agriculture has made impressive strides over the past 50 years in crop yield and input productivity growth, especially since the advent of genetically-modified crops in 1996. However, future growth rates could decline if U.S. agriculture does not sufficiently adapt to climate change. We examine the magnitudes of weather impacts on U.S. corn yields during 1960–2011—with a focus on intense precipitation and nitrogen use efficiency—and use the empirical results to forecast yields for the subsequent 20 years (2012-2031). We improve upon past methodologies by employing dynamic Bayesian regressions. These dynamic models permit rapid updating of new information, consistent with both pronounced yield growth in recent years and agricultural adaptation to changing growing conditions. We find that corn yields will increase by 27–41% over 2011 yields in top-growing states, though yields will gradually decline in less-productive states where climate change impacts could be among the most harmful. Our forecasts are generally robust to the empirical specification and assumptions about the econometric disturbance term, and have similar out-of-sample performance. To the extent that increasingly intense rainfall could contribute to nitrogen and other nutrient leaching, farmers may need to adjust nutrient applications in response to changing production environments.

Keywords Corn yields · Climate change · Biotechnology · Intense precipitation · Nitrogen use efficiency · Bayesian dynamic models · Information updating

J.R. McFadden
Economic Research Service, U.S. Dept. of Agriculture, 1400 Independence Avenue SW,
Mail Stop 1800, Washington, DC 20250-0002, USA
e-mail: jonathan.mcfadden@ers.usda.gov

J.A. Miranowski (✉)
Department of Economics, Iowa State University, 382B Heady Hall, Ames,
IA 50011-1054, USA
e-mail: jmirski@iastate.edu

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_16

1 Introduction

U.S. agriculture has made impressive strides over the last 50 years in crop yield and productivity growth, and this growth has been even more impressive since the advent of biotechnology introduced in 1996. No crop has been more important than corn in the U.S. biotechnology era, although other crops have also realized significant productivity gains. The productivity impacts of biotechnology have been even more important in developing countries where farmers have less control of the production environment. An excellent review of the potential and contribution of genetically-modified crops can be found in Barrows et al. [2].

The economics of corn farming in the United States has been evolving since the latter half of the twentieth century. In 1960, the U.S. harvested roughly 4 billion bushels of corn from 71 million acres, with approximately 1.6 million tons of nitrogen and 29 million pounds of pesticides (active ingredients) applied to corn. In more recent years, the U.S. has harvested 12–13 billion bushels from 87 million acres, while nitrogen and pesticides applied to corn have increased to 5.6 million tons and 204 million pounds, respectively (NASS [37]; Fernandez–Cornejo et al. [13]). Average U.S. corn yields have tripled over this period, up from 55 bu/ac in 1960 to 140–170 bu/ac in modern times (NASS [37]). The current dynamic context of US corn farming is one of increasing yields and increasing nitrogen use efficiency amidst stable- or slowly-increasing acreage allocations. These substantial changes are largely the result of initial development and later widespread adoption of biotechnologies and new information technologies.

Since the commercial introduction of Bt corn (using the soil bacterium, *Bacillus thuringiensis*) in 1996, research and development (R&D) of genetically-engineered seeds has brought significant development of new traits and multiple trait stacking. As of 2013, 76 percent of U.S. corn farmers planted Bt corn, and 85 percent planted herbicide-tolerant (HT) varieties, permitting more effective and less-costly weed control (Fernandez–Cornejo et al. [14]). Rapid farm-level adoption reflects that biotechnologies: (i) improve marginal productivities of several agricultural inputs, (ii) alter the optimal combinations of inputs and natural resources, and (iii) increase farm profitability. More efficient uptake and use of water and nitrogen by genetically-engineered corn plants have boosted nitrogen use efficiency and land productivity. Biotechnology, coupled with changes in other management practices such as decreases in row spacing and higher seeding rates, have contributed directly to higher yields. In turn, this has largely obviated the need to expand cropland acreage and significantly expand nitrogen applications to keep pace with rising food demand. As output expands and optimal input use remains relatively unchanged or declines, variable profits have slowly risen (Fernandez–Cornejo and Wechsler [12]). Although the general equilibrium effects are difficult to disentangle, aggregate adoption of biotechnology and related inputs will likely increase input productivity and market incentives that drive further R&D investment in the industry.

The advent of new information and data-based technologies are also complementing yield growth in the last decade. As with biotechnology, they influence marginal productivities of other production inputs, thereby changing the optimal input mix and ultimately farm output. For example, yield monitors are used to inform annual management decisions by improving nutrient use, pest control, energy, and operating efficiency. In the longer run, investments in irrigation, drainage, and capital equipment, ensure more accurate and efficient in-field operations. Variable rate applicators and guidance systems ensure optimal quantity and placement of seeds, fertilizer, and chemicals and reduce labor, energy, and machinery costs. Advances in information technologies complement biotechnologies and enhance marginal productivities of all inputs. As R&D expands and input costs decline, agriculture will benefit from continued use of bio- and information technologies, especially under changing climate conditions.

While bio-and information-technology developments are improving, yield and productivity growth may be partially or totally offset by climate change (Schlenker and Roberts [46]; Lobell et al. [25]). Accompanying corn yield and productivity growth has been fundamental and often adverse changes in weather inputs [32]. Much of the U.S. experienced an average temperature increase of roughly 1 °C or higher during 1901–2012. Annual precipitation during 1950–2010 increased by 5–25 mm/year per decade in the central U.S. In the absence of near-term climate change mitigation, projections indicate higher frequencies of extreme heat events, more intense droughts, greater precipitation variability, fewer frost days, and increases in heavy precipitation events (Romero–Lankao et al. [43]). Although there is large spatial variability of climate change impacts (and variation in confidence among climate change models), it is becoming clearer that corn farming in the highly-productive, central U.S. will need to adapt to an evolving production environment. The full range of climate adaptation strategies is unknown and cannot be forecast, but agricultural adaptation is not a new phenomenon. Agricultural technology and best management practices have been continually adapting and advancing over much of the last century (Schimmelpfennig et al. [45]; Zilberman et al. [54]). Continued adaption will likely involve R&D by bio- and information-technology firms and public research institutions (Heisey and Day-Rubinstein [18]), and marginal adjustments of inputs and cropping patterns at the farm level (Marshall et al. [27]). Both input and output adjustments on farms and R&D adjustments in industry and research institutions will also be influenced by agricultural policy as it adjusts to changing climatic and production environments.

Corn production, and corn yields in particular, result from significant interdependence between climate, information technologies, and biotechnologies in the near term. To capture the main aspects of dynamic corn production, we estimate Bayesian dynamic regressions using temporal variation in yields and weather over 1960–2011. Our Bayesian dynamic regressions improve on more conventional agricultural econometric methods by directly modeling outliers, structural change, yield skewness, and limited information content of older data. Our research examines four hypotheses: (i) corn yield growth has occurred over the past half-century despite climate change, (ii) future climate change impacts on yields will exhibit substantial state and regional variability, (iii) extreme weather, especially intense precipitation, has had negative effects on yields, and (iv) precipitation and nitrogen interact in determining yields.

We fit two competing regression specifications: one that incorporates average temperature and precipitation, and another that incorporates weather extremes. Coefficients are used to forecast yields in 2012–2031 for the 11 highest corn-producing states. We compare forecasts across three yield distribution assumptions: normal (using least squares), Student's *t*, and beta.

The balance of the paper proceeds as follows. Section 2 examines the related literature on climate change, bio- and information technologies, and corn yields. Section 3 lays out the econometric model and two estimation procedures. Section 4 presents both regression specifications and examines the in-sample and out-of-sample data. Section 5 provides our empirical results. Section 6 discusses the results and conjectures about the future of US corn yields, and Sect. 7 concludes. All tables and figures of results are contained in the appendix.

2 Corn Yields, Biotechnologies, and Climate Change

The present research stems from McFadden and Miranowski [29]. Using Bayesian dynamic models for the top 11 producing states, we found that Corn Belt corn yields will grow by 28–33% through 2031, with increases in Great Lakes states' yields up to 37%. Yield growth in the Great Plains states is less pronounced, and in certain less-productive areas, yields may decline. In states more suited to growing corn, nitrogen applications can partially mitigate harmful impacts of heavy precipitation. Regarding model fit, we found that regression models with *t*-distributed and beta-distributed yields have similar in-sample performance. McFadden and Miranowski [30] had a similar focus. Using data on 770 rainfed counties in 14 states, we confirmed that yields will increase by 10–40% over the next two decades. Long differences in weather variables are used to identify a long-run, cross-sectional relationship between climate change and technical progress in yields. After controlling for regional soil productivity and other possible confounds, we found that technical change in yields responds endogenously to climate change in the long run.

There is a large literature on the agricultural economics of bio- and information technologies. U.S. farm-level adoption of biotech corn has been rapid, driven by expectations of higher yields and input savings (Fernandez–Cornejo and Caswell [11]). This has led to changes in the nature of seeds research and the mix of public and private breeding research, with market structure implications for corn and soybeans (Foltz et al. [15]; Shi [47]; Shi et al. [48]; Huffman [23]). The increased potential for much higher yields has also expanded domestic biofuels markets and helped initiate advanced biofuels, though there is uncertainty surrounding biofuels policy (de Gorter and Just [8]; Miranowski [36] Rosburg et al. [44]). Consumer acceptance of biotech foods has been increasing in recent years (with notable exceptions among some market groups), with information effects and labeling having important roles in shaping consumer willingness-to-pay (Huffman et al. [22]; McFadden and Huffman [28]; McFadden and Huffman [33]).

Much less is known empirically about the range of adaptation mechanisms available to farms and their optimal uses under adverse weather conditions. In principle, farmers may move spring plantings forward to take advantage of warmer early-season temperatures and avoid late-season extreme heat (Ortiz–Bobea and Just [39]). The effectiveness of this strategy, however, could be limited by greater variability in frost days and distributional shifts in precipitation across the growing season.

Marginal adjustments in the timing and volume of irrigation water applications, where available, is another plausible adjustment strategy. Hornbeck and Keskin [21] show that farmers in the Great Plains substituted toward more water-intensive crops as irrigation water from the Ogallala Aquifer became available. In a similar study, Hendricks and Peterson [19] estimate that the demand elasticity for irrigation water in this region is very inelastic, so marginal adjustments in the near term seem likely. Over the long run, dwindling aquifer recharge rates could shift irrigated corn production to regions with more sustainable irrigation or rainfed regions (Marshall et al. [27]).

McFadden and Miranowski [31] address intensive and extensive margin adaptation to climate using data from several thousand farms in the central U.S. We find that early- and late-season temperatures and rainfall influence the selection or choice to grow corn. Mid-season weather patterns influence the choice of growing soybeans. Yet, crop switching, which may occur under mild climate change scenarios, is far more influenced by soil productivity. In other words, the choice of growing corn, soybeans, and other crops is driven more by soil productivity factors than mild climate change adjustments.

Our research fits more generally in the economics of agricultural yields. These studies vary with respect to forecasting, in-sample weather impacts, and distributional form. Miranowski et al. [35] use data on the top 17 corn-producing states during 1960–2011 for forecasting. There is evidence of at least one structural break in each state, and linear trend and autoregressive models are estimated around these breaks. Corn yield forecasts indicate increases of 1–4 bu/ac per year through 2030, depending on the presence of short- or long-run trends.

Regarding weather and yields, several studies underscore the importance of changing weather patterns and biotech adoption rates for in-sample studies (Schlenker and Roberts [46]; Roberts et al. [42]; and Xu et al. [52]). The emphasis has been mainly on temperature-related variables (e.g., growing degree days) and drought indicators. More statistical approaches have been implemented by Harri et al. [17] and Claassen and Just [4]. These studies find evidence of yield skewness and other factors contributing to non-normality that vary over growing regions. The most suitable distributions for modeling yields remains an open question.

3 Econometric Model

We begin by estimating all regression models with ordinary least squares (OLS). These estimates provide a useful benchmark for comparisons among the two dynamic models explained below. Assuming normally-distributed yields, differences between least squares and the dynamic t-distribution results primarily reflect the differences between static and dynamic estimation methods.

3.1 Dynamic Bayesian Regressions

The first dynamic framework is a linear state space model. State space models have an observation equation and a state equation. The dependent variable at any time period is a linear function of unobserved states and a random disturbance. The law of motion for unobserved states is a random walk. We estimate the following system:

$$Y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim N(0, k_t \phi_t^{-1}) \tag{1}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t \theta_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim t_{n_{t-1}}(\mathbf{0}, \mathbf{W}_t). \tag{2}$$

In period t, Y_t is the dependent variable, $\mathbf{F}_t^T$ is a $(1 \times n)$ vector of regressors, $\boldsymbol{\theta}_t$ is an $(n \times 1)$ vector of regression coefficients (state parameters), $\mathbf{G}_t$ is the system evolution matrix, ν_t is the observation error, and $\boldsymbol{\omega}_t$ is the system disturbance vector. Note that (1) is the observation equation, and (2) is the state evolution equation. Error terms satisfy temporal and mutual independence: $Cov(\nu_s, \nu_t) = 0, Cov(\boldsymbol{\omega}_s, \boldsymbol{\omega}_t) = \mathbf{0}_{n \times n}$ for all $t \neq s$, and $Cov(\nu_s, \boldsymbol{\omega}_t) = \mathbf{0}_n$ for all t, s. The observation variance is the product of a known variance dispersion parameter, k_t, and ϕ_t, the observation's precision, which has a gamma prior distribution. The system disturbance is from a mean-zero, multivariate t-distribution with degrees of freedom that are updated sequentially and a block-diagonal variance (scale matrix), $\mathbf{W}_t$. The three submatrices comprising $\mathbf{W}_t$ are an intercept block, a regression block, and a time trend block. Explanatory information decays at different rates, so each of the blocks is adjusted by a separate discount factor: δ_{int}, δ_R, and δ_{tr}, respectively.

Priors on coefficients and the observation variance are the following:

$$\theta_t | I_{t-1} \sim t_{\delta_t n_{t-1}}(\boldsymbol{\alpha_t}, \mathbf{R}_t) \tag{3}$$

$$\phi_t | I_{t-1} \sim \Gamma\left(\frac{\delta_t n_{t-1}}{2}, \frac{\delta_t d_{t-1}}{2}\right), \tag{4}$$

where $\boldsymbol{\alpha_t}$ and $\mathbf{R}_t$ are the location and scale parameters, respectively, of the multivariate t-distribution with $\delta_t n_{t-1}$ degrees of freedom. The shape and scale parameters of the gamma distribution are $(\delta_t n_{t-1}/2, \delta_t d_{t-1}/2)$, respectively. Knowledge available at time $t - 1$ is contained in the information set, I_{t-1}. The discount factor here, δ_t, is a general discount and is set very close to unity, e.g., 0.99.

The posterior distributions and one-step ahead forecasts are:

$$\boldsymbol{\theta}_t | I_t \sim t_{n_t}(\mathbf{m_t}, \mathbf{C}_t) \tag{5}$$

$$\phi_t | I_t \sim \Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right) \tag{6}$$

$$Y_t | I_{t-1} \sim t_{\delta_t n_{t-1}}(f_t, Q_t). \tag{7}$$

Although the dependent variable depends linearly on regression coefficients, the system of recursive equations used for estimation involve several nonlinearities affecting posterior distributions. Note that k-step ahead forecast distributions are updated similarly. See Pole et al. [41], West and Harrison [51], and Durbin and Koopman [9] for more details.

3.2 Time-Varying Beta Distributions

The thick-tailed t-distributions in dynamic linear models (DLMs) are beneficial for modeling outliers, but one drawback is their symmetry. For several U.S. states, high yields occur frequently and low yields infrequently, suggesting that conditional distributions should be negatively skewed.[1] The beta distribution permits flexibility in modeling skewed yields (Day [7]; Claassen and Just [4]). The support of the beta distribution is [0, 1]. We use a four-parameter transformation in which each state's yield at time t is $(y_t - \max)/(\max - \min)$, where max and min are state-specific maximum and minimum obtainable yields.[2]

The time-varying beta model builds on da Silva et al. [6] and Lopes and Tsay [26]:

$$Y_t|\mu_t \sim Beta(\phi\mu_t, \phi(1-\mu_t)) \tag{8}$$

$$\mu_t = (1+\exp[-\boldsymbol{\beta}_t\mathbf{x}_t])^{-1} \tag{9}$$

$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \mathbf{W} \sim N(\boldsymbol{\beta}_{t-1}, \mathbf{W}). \tag{10}$$

The system moves according to (10), in which the states, $\boldsymbol{\beta}_t$, are given independent normal distributions with means $\boldsymbol{\beta}_{t-1}$. The beta distribution is a member of the exponential family of distributions, and (9) links the states and regressors, $\mathbf{x}_t$, to the dependent variable's mean, μ_t. The time-invariant precision parameter, ϕ, is inversely related to the dependent variable's variance, $\mu_t(1-\mu_t)/(1+\phi)$.

We use the Liu and West [24] particle filter to estimate (8)–(10). Readers interested in the exact details of this algorithm or similar sequential Markov Chain Monte Carlo methods are referred to Durbin and Koopman [9], Liu and West [24], and Migon et al. [34]. We first specify priors: $\phi \sim Inv\Gamma(\alpha_p, \beta_p)$, and distinct $W \sim Inv\Gamma(\alpha_w, \beta_w)$

[1] We analyze quantile–quantile (QQ) plots of OLS residuals for various regression specifications across U.S. states. The plots indicate substantial skewness for certain states. We also undertake Shapiro–Wilk tests of normality and Kolmogorov–Smirnov tests that the errors come from a beta (5, 2) distribution. A skewed distribution is more appropriate for some states, but we cannot reject the null hypothesis of normality for five states.

[2] Several techniques can be used to estimate the four parameters of the reparameterized beta distribution. We impose minimum and maximum yields proportional to each state's observed minimum and maximum. Specifically, maximums are set at 150% of observed, state-specific maximums, while minimums are set at five bushels fewer than the smallest observed yields.

and $\beta_0 \sim N(m_0, C_0)$ priors for each entry of the initial $\boldsymbol{\beta}$. The hyperparameters are $m_0 = 0$, $\alpha_p = 20$, $\beta_p = 315$, and $C_0 = 0.3$. For W corresponding to the first component of $\boldsymbol{\beta}$, $\alpha_w = 3$ and $\beta_w = 0.2$. For all other entries of $\boldsymbol{\beta}$, we increase α_w by 0.1 with corresponding $\beta_w = (\alpha_w + 1)0.05$. There is a tuning parameter in the model, which is set to 0.97 and consistent with a discount factor of roughly 0.95. To forecast, we first simulate the states from (10), taking medians across all replications of sequences. These simulated values are then inserted into (9), as well as the weather regressors. The last available estimate of ϕ at $t = 2011$ is used for all forecasts.

4 Regression Specifications and Data

Prior to estimation of both regression specifications given below, we estimate a specification containing only an intercept and linear time trend. They are denoted below as “trend only”. This is designed to illustrate potential forecast biases from models that do not account for a crucial component of agricultural production (e.g., exogenous weather).

4.1 Baseline Specification

Agricultural economics and agronomy suggest a basic regression model that uses temperature and precipitation variables during important periods of the growing season to explain yields. To identify the effects during important growth stages, we use statewide means of precipitation and temperature for the months of May, June, July, and August. Our sample for 1960–2011 is composed of the top 11 corn-producing states for 2011 (producing 2% or more of U.S. production). Gridded monthly averages of daily data are from the National Oceanic and Atmospheric Administration’s (NOAA) U.S. Climate Divisional Database. A linear time trend is used to detrend yields. For the j^{th} state in period t, the baseline specification is:

$$\mathbf{F}_{j,t}^{T} = \left(1 \; MonthlyTemp_{j,t} \; MonthlyPrec_{j,t} \; (JulyPrec_{j,t})(N_{j,t}) \; T\right). \quad (11)$$

The $(JulyPrec_{j,t})(N_{j,t})$ term captures the interaction of July rainfall and nitrogen.[3] After estimation, highest posterior density (HPD)-based tests help reduce the model to a subset of important variables. The statistic is compared at each time point

[3] Nitrogen use data are from the Economic Research Service (ERS) of the United States Department of Agriculture (USDA). There are missing data in all states for certain years. We use zero-intercept regressions of state-level nitrogen use on total U.S. nitrogen use to impute the missing values. The nitrogen categories included are anhydrous ammonia, ammonium nitrate, ammonium sulfate, nitrogen solutions, and urea. Nitrogen forecasts for 2013–2031 are obtained by OLS regression on time.

to the 95% critical value of the F distribution. If the critical value is exceeded for approximately 30% of the 52 years in our sample, we include the regressor.[4]

4.2 Alternative Specification

A drawback of the baseline model is its simplification of more complex biological relationships, especially regarding extreme weather. To incorporate some of these effects, researchers have proposed Growing Degree Days (GDD), Heating Degree Days (HDD), Extreme Degree Days (EDD), Killing Degree Days (KDD), and other transformations of temperatures that account for accumulated beneficial or extreme sunlight. In addition, diurnal temperature range (DTR), the difference between the daily maximum and minimum temperatures, is used to measure overnight plant cooling.

For transparency, we capture temperature extremes by indicators of a monthly average maximum temperature equaling or exceeding 90 °F in July and August (denoted below as JulH and AugH).[5] In several sample states, the corn plant begins to pollinate and develop kernels during July and August growth stages that are sensitive to weather extremes. The two dummy variables rely on data from NOAA's Global Historical Climatology Network (GHCN), which are station-level data of monthly means from daily maximum and minimum temperatures. We also use July and August DTR, with data again from GHCN.

High rainfall rates can lower yields by leaching plant nutrients in soils. Rainfall runoff and erosion can transport nutrients on and bound to the soil surface and limit fertilizer effectiveness. Rainfall intensity is measured by counts from weather stations receiving at least one inch of rain per hour, similar to daily heavy precipitation events previously tracked by EPA [10]. Counts of hourly rainfall events are constructed from hourly, station-level data in NOAA's Cooperative Observer Network (COOP).[6] There are very few sources providing forecasts of extreme rainfall events (Groisman et al. [16]). We use negative binomial regressions on time to generate forecasts.[7] The alternative specification for the j^{th} state in period t is:

[4] This relative frequency criterion is robust to selecting other threshold values. Individual t- and joint tests from state-level OLS estimation agree with our HPD-based tests for a majority of regressors across states.

[5] This cutoff is a compromise between the 86 °F used in canonical GDD formulas and Lobell et al. [25], the 84.2 °F suggested by Schlenker and Roberts [46], and the 90 °F of Xu et al. [52]'s excess heat degree days.

[6] No cooperative has a complete data record. Missing data could be imputed based on neighboring data, but this could worsen measurement error. Heavy rainfall events occur during thunderstorms, which NOAA [38] suggests average 15 miles in diameter with 30-minute lengths. The typical closest station is 5–10 miles away from the observing station, where rainfall rates could significantly differ.

[7] Likelihood ratio tests reject the null hypothesis of equidispersion, assumed in Poisson regression models. Estimates from the Poisson and negative binomial regressions are very similar and do not alter the forecasts.

$$\mathbf{F}_{j,t}^{T} = \big(1\ JulyDTR_{j,t}\ AugDTR_{j,t}\ JulyH_{j,t}\ AugH_{j,t}\\ RainEvent_{j,t}\ (JulyEvent_{j,t})(N_{j,t})\ T\big)\,, \tag{12}$$

where $RainEvent_{j,t}$ are separate counts of heavy rainfall events for May, June, July, and August. We do not develop a reduced model. Distinct from the baseline model, the nitrogen interaction uses July rainfall events.

4.3 Climate Change Scenarios

Given the evidence that current weather conditions reflects global warming, future weather patterns will not be similar to those during 1960–2011. Using the in-sample weather variables for forecasting is thus unsuitable. One common practice is to assimilate output from multiple realizations of one or more global climate models (GCMs). There are drawbacks (e.g., differing GCM assumptions and conflicting GCM results), but this practice is reasonable for our purposes (Auffhammer et al. [1]).

Data on precipitation, average surface temperatures, and minimum and maximum daily temperatures for the next two decades (2012–2031) are taken from the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project phase 5 (CMIP5), available from the Bureau of Reclamation, Department of the Interior [3]. These GCMs have been downscaled to 1/8° resolutions and adjusted for mismatches between simulations and the historical record. The output is classified according to four "levels" of climate change.[8] Results in the next section only consider mild climate change, i.e., RCP 2.6. Estimates and forecasts under severe climate change, RCP 8.5, are very similar to the RCP 2.6 results. This is because significant departures from current weather patterns are not projected for the next two decades.

Climate projection data are available by latitude and longitude. To ensure that the projections are not influenced by a few GCMs or specific runs of a particular GCM, we average over several models, some with multiple runs.[9] Our choice of models is guided by Pierce et al. [40], which gives a ranking of "high skill" models.

[8] These four levels are representative concentration pathways (RCP) 2.6, 4.5, 6.0, and 8.5. The pathways index radiative forcing, the rate of change in the difference between incoming and outgoing solar energy in the atmosphere. Larger radiative forcing indicates more severe climate change.

[9] Averages are taken over 13 GCMs: CanESM2, CCSM4, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-R, HadGEM2-AO, HadGEM2-ES, MIROC-ESM, MIROC-ESM-CHEM, MIROC5, MPI-ESM-MR, and MPI-ESM-LR (Reclamation [3]).

4.4 Descriptive Statistics

Table 1 contains descriptive statistics for the baseline model in our 11-state sample for 1960–2011. Early-season temperatures range in 54–64 °F, while temperatures in June, July, and August are in 64–79 °F. The hottest month, July, is also the month with the lowest variability in temperatures. Similarly, May experiences the lowest average temperatures but exhibits the largest standard deviations. On average, the warmest state is Kansas, with monthly average temperatures near 80 °F. Michigan is the coldest state but exhibits the most stability in (relatively low) average rainfall, 3.1–3.4 in. per month in the growing season. The traditional Corn Belt states of Iowa, Illinois, and Indiana have abundant monthly rainfall, 3.6–4.8 in. South Dakota is a perennially dry state, averaging roughly 2 inches in August.

Descriptive statistics for regressors in the alternative specification are given in Table 2. The Corn Belt states of Iowa, Illinois, and Indiana have had the lowest average monthly DTR, 21–22 °F. This contrasts with warmer and drier states, such as South Dakota and Kansas, which have the largest DTR. Both of these states also have many occurrences of hot monthly average maximum temperatures in July and August. 85% of the July months in Kansas during 1960–2011 have experienced average maximum temperatures exceeding 90 °F. Only 17% of July months for Illinois and 48% for Nebraska have had similar temperature effects. Missouri, Iowa, and Illinois have relatively high counts of intense hourly rainfall, while Michigan and South Dakota have lower instances of intense hourly rainfall. This pattern is evidenced in the average precipitation statistics of Table 1. Intense rainfall is concentrated in June and July, months in which there are higher frequencies of storms and thunderstorms.

5 Empirical Results

Figure 1 depicts forecast means and 95% credible intervals over 2012–2031 in the baseline dynamic linear model for Iowa and Illinois.[10] Both states begin in 2012 at very similar points, 160–170 bu/ac, but evolve differently over the forecasting window. Means increase roughly linearly over the 20 years by 2–5 bu/ac for Iowa and 1–8 bu/ac for Illinois. The mean gap between states is small for all years and does not exceed 15–20 bu/ac, which replicates the pattern observed in most years in the data. This is a consequence of similar weather, soil productivity, management techniques, and cropping patterns for Iowa and Illinois in recent years. Iowa's forecasted 2031 mean is 219 bu/ac, a 27% increase over 2011 yields, whereas Illinois' mean is 206 bu/ac, a 31% increase over 2011 yields. Robust yield growth is similar for other Corn Belt states in our study and compares well with the county-level results in other studies (Xu et al. [52]; McFadden and Miranowski [29]). The most noticeable difference between forecasts is the substantial uncertainty in Illinois yield growth.

[10] For space considerations, we have narrowed the focus of our discussion. Results for all states are available upon request. Similar state-level results are obtained in McFadden and Miranowski [30].

The 95% credible region for Iowa yields in 2031 is [185, 250] bu/ac, but we cannot rule out slight declines in Illinois yields over the next decades.[11]

We next turn to a comparison of forecasts for two Great Lakes states, Minnesota and Wisconsin, in Fig. 2. Minnesota and Wisconsin are interesting to consider because of similarities in weather but large discrepancies in soil productivity and management practices. As with the Corn Belt states, Minnesota and Wisconsin begin similarly at 170 bu/ac but then have overlapping forecast means throughout the remainder of the two decades. In 2031, forecast means are 217–220 bu/ac, representing 41% and 39% increases on 2011 yields in Minnesota and Wisconsin, respectively. Unlike estimates for other regions, forecasts for Great Lakes states are smoother and more linear. This is a consequence of smooth climate model projections for average temperature and precipitation in more northern latitudes. Moreover, Great Lakes states have experienced milder temperatures and less variable rainfall historically, which is reflected in regression slopes that permit greater yield growth. However, we forecast much higher yield variability for Wisconsin over the next several years, with equal possibilities of greatly-increasing or slightly-declining yields. The relatively large variances for Illinois and Wisconsin could be partially driven by lake-effect weather patterns associated with proximity to Lakes Superior and Michigan.

The forecasts for Nebraska and Kansas in Fig. 3 stand out from those of other highly-productive states for several reasons. Nebraska's yield growth is similar to that of Iowa and Minnesota. Average yields increase 38% from the 2011 level of 160 bu/ac to the 2031 level of 221 bu/ac. Among Corn Belt states, Nebraska exhibits much less forecast uncertainty, with narrow credible regions that extend 15–20 bu/ac about the mean. Corn production in western Nebraska relies on aquifer irrigation, while crops in eastern Nebraska are either rainfed or rainfed with supplemental irrigation from nearby river water. In recent years, Nebraska yields have been more stable than in other Corn Belt states, in part because of unique weather and sufficient supplies of groundwater irrigation. This is in sharp contrast to Kansas. Our forecasts indicate a gradual decrease in yields over the coming years but with much year-to-year variation. In the last three years of the sample, Kansas yields declined 31% from 155 bu/ac in 2009 to 107 bu/ac in 2011. This marked yield decrease is incorporated in our models' dynamic updating and provides forecasts that are less optimistic than static models, e.g., least squares. To the extent that adverse weather and dwindling irrigation inputs partially caused the recent yield decline and will persist or increase in the near future, the dynamic forecast means will perform better than the least squares forecasts.[12]

To illustrate the model dynamics underlying our baseline forecasts, estimated coefficients for the July rainfall-nitrogen interaction in Iowa are provided in Fig. 4. We choose to illustrate coefficients for the time-varying beta model because the

[11] The model-selected discount rate for Illinois is 0.90, which partially contributes to a large variance. In general, the short-run yield-growth trend in Illinois is larger than its long-run average trend, in agreement with the Illinois results in Miranowski et al. [35].

[12] Forecasts for several rainfed counties in Kansas are also largely declining through 2031 (McFadden and Miranowski [29]). This suggests that much warmer and drier weather is an important concern for areas that cannot use irrigation for mitigation.

scaling is more straightforward to interpret. The figure shows that rainfall-nitrogen interactions have evolved stably over time. At the beginning of our sample, the interaction effects first increase and then moderately decrease, but there are no abrupt breaks in the movement of the estimates. This supports our use of dynamically-updated models for forecasting. The marginal impacts of nitrogen, evaluated at the in-sample July rainfall data, have nonlinear effects. This is because the marginal impact is a nonlinear (logistic) function of variable July rainfall and all other weather regressors. Averaged over 1960–2011, the marginal impact of the nitrogen interaction is 0.53 bu/ac. Nitrogen availability generally enhances the marginal productivity of July precipitation, which improves pollination and corn ear fruiting. In other words, adequate July precipitation improves the marginal productivity of nitrogen applied in the growing season.

Evolution in the July rainfall event-nitrogen interaction from Iowa's alternative specification, shown in Fig. 5, differs from that in the baseline specification for several key reasons. Noticeable jumps in the coefficients occur over several years. There is a downward trend during 1967–1969 and then a distinct one-time drop in 1983, consistent with evidence of breaks in past research (McFadden and Miranowski [30]; Miranowski et al. [35]).[13] Given the widespread drought in mid-1983, a substantial drop in coefficients involving summer precipitation is intuitive and expected. There is also a pronounced downward trend near the sample endpoint, with a modest uptick in 2011. This is the result of improved nitrogen use efficiency in recent years and a relatively lower number of intense hourly rain events in Iowa during July 2011. Similar to the baseline estimates, the interaction effect has an average positive marginal impact on yields. However, our ability to give more precise conclusions is limited by the wider 95% credible intervals.

Model performance is assessed in Table 3 using a mean absolute deviation (MAD) criterion. We average the absolute value of the difference in forecasted and actual yields over 2012–2014. These are calculated for the three models (OLS, dynamic t-distribution, and dynamic beta distribution) and three specifications (trend-only, baseline, and alternative regressors). For Iowa, Nebraska, and Minnesota, three of the top four producing states, the OLS model with the baseline regressors has the highest out-of-sample forecast accuracy. These three states experienced prolonged drought and episodes of high heat during 2012 and 2013 that reduced yields from recent upward trends. However, most models perform similarly well, especially the t-distributions for Iowa and Nebraska and the beta distributions for Minnesota. The difference between these forecasts and those of the best models is 1–4 bu/ac.

Several other interesting features can be inferred from Table 3. First, practical differences in forecast accuracy between static and dynamic models are illustrated in several states. For example, the best performing model for Illinois is a trend-only dynamic t-distribution, though none of the models are highly accurate. Illinois 2014 yields reached a record high of 200 bu/ac, roughly 27% over its 157 bu/ac average in 2011 and a 90% increase on its 2012 yields of 105 bu/ac. This validates a basic tenet of

[13] Bayesian model monitoring based on cumulative sum (CUSUM) techniques suggest a cumulative breakdown of model fit in several years, including 1983.

forecasting: no plausible models can reliably forecast highly-erratic time series over long time periods. However, our dynamic Bayesian methods perform better because of the underlying updating procedures. This is also confirmed for Michigan, where yields moved slowly upward over 2007–2011. Static methods underestimate more recent yield growth because of flatter regression slopes needed to accommodate early, less-informative data. Second, models that account for weather extremes forecast better in states where these have occurred more frequently and are more likely to occur in the future. Forecasts from the alternative beta distribution matched poor yields in Kansas in 2012 and 2013 much better than the optimistic OLS forecasts. In South Dakota, a relatively warm and dry state, the best model is the alternative dynamic *t*-distribution. Note that for these states, temperatures above 90 °F tend to decrease yields, but intense hourly precipitation can boost yields in particularly dry years. Third, several highly-productive regions with milder weather do best under the baseline beta-distributed model. This is intuitive for the eastern Corn Belt states of Indiana and Ohio. Ample rainfall and cooler growing seasons point to the baseline beta model as a reasonable choice and is confirmed by the lowest deviations, 31.3 and 13.7 bu/ac, respectively.

In sum, our results indicate significant yield growth over the past half century and into the twenty-first century in the face of substantial climate change. Although intense precipitation reduces yields, mean precipitation and nitrogen applications interact in raising or lowering yields, depending on the limiting factor. A critical dimension of our analysis is the geographic variability in climate change impacts, as indicated by our state-by-state comparisons in Figs. 1, 2 and 3. Climate change will have important regional impacts based on soil productivity, management practices, and geographical features contributing to distinct weather patterns. Our state-level results are an informative approach for comparing impact magnitudes among neighboring (or otherwise similar) states. In this sense, the results are in broad agreement with economic theory emphasizing the geographic distribution of impacts (Zilberman et al. [53]).

6 Discussion

Our dynamic estimates and forecasts have several policy and adaptation implications for an increasingly-dynamic production environment. A crucial feature of rising yields has been soil productivity and conservation. Successful adaptation to a changing climate requires improved information about soil management practices in changing production environments. Additionally, we should recognize important soil productivity limitations in a changed climate. Our state-specific regressions incorporate soil productivity, and productivity of other inputs, indirectly in the intercepts. Although soil organic matter may fluctuate slightly across years depending on cropping choices and practices, inherent soil productivity is largely time-invariant if sound soil conservation is practiced. This could restrict the potential to shift corn production into regions with more advantageous climates, and at the same time sustain higher yields if inherent soil productivity is limited, i.e., soil is only productive if timely moisture and favorable temperatures are available. In other words, soils and weather are both limiting factors in production. This generates an important soil productivity-climate tradeoff that should be considered in farmers' dynamic economic decision making (McFadden and Miranowski [31]). In upcoming decades, Minnesota, Michigan, and Wisconsin will likely experience a comparative advantage in beneficial climate, but this will be offset by poorer soil productivity in many areas or states, particularly in Wisconsin and Michigan. More generally, the sandier soils of the Great Lakes regions and the clay-dominant soils of the Southern Plains are not ideal soil structures for commercial corn growing. To the extent that our models are capturing poorer soil productivity with smaller intercept coefficients, production shifts to more northern latitude states are viable adaptation options but not viable economic options.

The dynamic interdependence between climate change and irrigation water availability is another crucial component of our analysis. Withdrawal rates of Ogallala Aquifer water have exceeded recharge rates in some areas south of Nebraska. In response to rising water costs and reduced availability, many irrigated operations in this region have gradually curtailed or ceased irrigation water applications. Given the lack of competitive sources of water, this outcome likely contributes to sizeable yield forecast divergence for Nebraska and Kansas in Fig. 3. The extent to which advanced information systems will alter the relationship between irrigation, weather, and other inputs is a similarly important issue.

The dynamics of corn production in recent years have been influenced by innovative adaptation strategies that rely on more computer-based monitoring and measurement data and management systems. In an era of knowledge technology and big data systems, many agricultural supply industries are working together to develop knowledge-based systems to facilitate farmers' adaptation of crop management decisions to climate change. These firms are developing climate-focused technologies for farmers, collecting extensive production data, monitoring weather and growing conditions, and providing real-time management recommendations on fertility, planting, pest control, and harvest activities. At the core of these technologies are networks of sensors and monitors providing near-time weather forecasts and improved information to farmers. Pending government approval, unmanned aerial vehicles will be used for scouting and monitoring nutrient, pest, and crop conditions and directing applications where needed.

Other closely-related adaptation strategies include adoption of precision and prescription agricultural innovations. Increasing seed industry concentration has boosted R&D for seed-based technologies with substantial weather and climate interactions. Field trials for drought–tolerant corn indicate significant opportunities for limiting downside risk of drought and heat spells. One source of complication, though, is a yield penalty incurred under excessive rainfall in possible inundation conditions, especially during certain stages of plant growth. Modern plant breeding is using bioinformatics to rapidly select for traits from existing corn seed lines to design plants appropriate to field conditions, i.e., and developing prescriptions to optimize yields. Combined with wireless communication, GPS guidance, and variable rate applicators, the result is a more automated system that increases yields, lowers energy and seed costs, and reduces the likelihood of pest infestations. However, adoption of these technologies may depend to some extent on the evolution of agricultural policy.

U.S. government policy has, at times, significantly influenced the price, location, crop mix, and acreage of agricultural production, especially in corn. USDA has been looking into the dynamic relationships between climate change and corn production for many decades. Impact assessments and steps toward greater climate change preparedness are summarized in its national climate change adaptation plan (USDA [50]). Conservation policies may need revision over space and time as intensity and quantity of rainfall and temperature change. Given increased potential for soil erosion from intensifying precipitation, conservation practices may need to be revised from using the universal soil loss (USLE2) base to reducing peak soil loss. From a water quality standpoint, intense rainfall events may increase nutrients being flushed from tile drainage and call for innovative drainage management strategies. EPA may have to revise point source pollution guidelines for animal feeding operations especially with increasing intense rainfall.

The 2014 Farm Act is shifting emphasis toward increased insurance and risk management strategies relative to traditional program payments and conservation incentives. Although risk is likely to increase with extreme weather, so is the need for conservation incentives to offset soil, nutrient, pest resistance, and productivity losses as well as water quality deterioration. USDA's Risk Management Agency (RMA) paid nearly $12 billion in losses from adverse weather in 2013. Under increasing likelihood of climate-induced losses, RMA's total annual indemnity payments may increase. Over the long run, innovations in insurance products and non-insurance risk management tools may be needed to limit large increases in weather-related risk and the associated budget exposure. This may entail partnerships between RMA and private firms to design more information-based and climate-focused risk reduction tools. There have already been recent innovations in climate-focused insurance products that are likely to continue as yield volatility increases with changing climate.

7 Conclusion

We have now entered a new era of knowledge-based information management systems. Biotechnology is an integral and necessary part of these knowledge-based systems which involve complementary information technologies, precision monitoring and farming practices, utilization of big data, bioinformatics in seed selection and development, prescription agriculture, and knowledge-based management systems to assist in adjusting and adapting to more extreme weather. Agronomists attribute yield growth to genetic potential of a variety, environmental factors, and crop management, as well as the interaction of these factors. Much of corn yield growth, roughly 60–80%, is attributed to improved genetics or crop breeding (Smith et al. [49]; Crosbie et al. [5]). Information and other knowledge technologies provide an opportunity to enhance the contribution of management to corn yield growth. As Smith et al. [49]) state, "More effective use of genetic diversity and crop management will allow U.S. maize breeders and farmers to accommodate climate change for the foreseeable future." What is happening in corn production parallels what occurred in poultry, pork, and dairy livestock production, as discussed in Hennessy et al. [20]. We are rapidly moving from "attentive husbandry or management" to "knowledge husbandry or management" by corn and other crop farmers. We are now developing knowledge-based technologies and management systems that improve input and output productivity in the face of climate change.

New knowledge-based firms, many tied to major seed and equipment firms, provide real-time management recommendations and other services to farmers. These knowledge-based management services frequently guarantee improved profitability or refunded subscription charges. Such information technologies and real-time management services, coupled with tailored seeds from modern plant breeding, monitoring, and variable-rate nutrient and pest control practices tailored to climate and field conditions, are truly the advent of prescription agriculture and may ensure continued yield growth in the face of climate change. The rapid adoption of biotechnology laid the cornerstone for significant yield increases almost two decades ago. In the past decade, modern information and plant breeding systems have helped sustain yields and productivity. Now, the advent of knowledge-based crop management systems further motivates our dynamic forecasts and is likely to bolster yield growth through mid-century. Further reductions in computational, monitoring, and management delivery costs in an era of high land and other prices, is likely to lead to continuing substitution of information and management technologies for more scarce inputs like labor, capital, nutrients, and land. These improved information-based knowledge systems may also provide improved options for monitoring and managing nutrients, soil loss, pest resistance, and other agricultural externalities.

Acknowledgements We thank participants at the 2014 Berkeley Bioeconomy Conference for helpful comments. This research was sponsored by the U.S. Dept. of Agriculture (USDA) Earth System Modeling program, Award 2013-67003-20642. The views expressed in this paper are those of the authors, and not necessarily those of the U.S. Department of Agriculture or the Economic Research Service. All errors are ours.

Appendix

Table 1 Summary statistics, 1960–2011. Baseline specification weather data. Table entries: mean (Std. Dev.)

State	May temp.	June temp.	July temp.	August temp.	May rain	June rain	July rain	August rain
IA	60.1	69.5	73.7	71.2	4.3	4.8	4.3	4.1
	(3.2)	(2.1)	(2.3)	(2.5)	(1.5)	(1.9)	(1.8)	(1.9)
IL	62.4	71.7	75.3	73.3	4.4	4.2	4.0	3.6
	(3.4)	(2.0)	(2.1)	(2.5)	(1.7)	(1.5)	(1.3)	(1.3)
NE	58.8	68.7	74.6	72.3	3.7	3.8	3.1	2.7
	(2.8)	(2.5)	(2.3)	(2.4)	(1.4)	(1.4)	(1.2)	(0.94)
MN	54.9	64.5	69.2	66.9	3.2	4.1	3.8	3.5
	(3.4)	(2.5)	(2.5)	(2.5)	(1.1)	(1.2)	(1.2)	(1.0)
IN	61.5	70.6	74.1	72.3	4.5	4.2	4.3	3.7
	(3.5)	(2.0)	(2.0)	(2.3)	(1.7)	(1.4)	(1.5)	(1.2)
SD	56.3	66.2	72.9	70.8	3.1	3.5	2.7	2.1
	(2.8)	(2.7)	(2.9)	(2.6)	(1.4)	(1.2)	(1.1)	(0.71)
WI	55.0	64.4	69.2	67.0	3.6	4.1	3.9	4.0
	(3.4)	(2.3)	(2.3)	(2.3)	(1.3)	(1.6)	(1.3)	(1.4)
OH	60.0	69.0	72.7	71.2	4.1	3.9	4.1	3.6
	(3.4)	(2.0)	(1.9)	(2.2)	(1.6)	(1.4)	(1.2)	(1.2)
KS	63.7	73.4	79.0	77.0	4.0	4.2	3.5	3.2
	(2.8)	(2.4)	(2.4)	(2.6)	(1.5)	(1.3)	(1.6)	(1.3)
MO	64.2	72.8	77.5	75.8	4.9	4.4	4.0	3.6
	(3.0)	(2.1)	(2.1)	(2.6)	(1.8)	(1.5)	(1.6)	(1.4)
MI	54.1	63.6	68.3	66.6	3.1	3.3	3.1	3.4
	(3.4)	(2.3)	(2.3)	(2.2)	(1.1)	(1.0)	(0.77)	(1.0)

Note Temperatures for May, June, July, and August are in °F. Rainfall for May, June, July, and August are in inches. Note that included regressors vary by state, depending on the results of Bayesian HPD-based tests. See text for data sources

Source NOAA U.S. Climate Divisional Database

Table 2 Summary statistics, 1960–2011. Alternative specification weather data. Table entries: mean (Std. Dev.)

State	July DTR	August DTR	July 90	August 90	May event	June event	July event	August event
IA	21.8	22.1	0.10	0.04	11.9	29.3	31.7	26.0
	(1.9)	(1.8)	(0.30)	(0.19)	(9.9)	(17.8)	(15.8)	(17.0)
IL	21.6	22.1	0.17	0.12	11.9	21.2	27.7	19.3
	(1.6)	(1.5)	(0.38)	(0.32)	(8.3)	(11.4)	(12.3)	(11.0)
NE	26.2	26.5	0.48	0.21	8.8	19.7	18.3	15.1
	(2.0)	(1.8)	(0.50)	(0.41)	(6.5)	(13.2)	(10.5)	(9.9)
MN	23.2	23.3	–	–	3.9	13.1	17.1	13.8
	(2.0)	(1.7)	–	–	(3.4)	(8.1)	(8.4)	(7.3)
IN	21.8	22.1	0.10	0.04	9.5	17.7	23.3	17.2
	(1.9)	(1.8)	(0.30)	(0.19)	(6.5)	(9.8)	(11.8)	(8.6)
SD	27.2	28.1	0.37	0.23	3.4	8.6	9.1	6.7
	(2.5)	(2.0)	(0.49)	(0.43)	(3.9)	(5.4)	(5.2)	(3.7)
WI	23.0	22.5	–	–	4.6	12.3	14.0	14.6
	(1.9)	(1.9)	–	–	(4.0)	(8.8)	(6.9)	(9.5)
OH	22.1	22.4	0.06	–	7.3	16.4	21.9	17.1
	(1.7)	(1.6)	(0.24)	–	(4.8)	(10.8)	(11.6)	(9.5)
KS	25.3	25.6	0.85	0.67	17.3	28.5	23.9	20.5
	(2.1)	(2.1)	(0.36)	(0.47)	(9.2)	(12.3)	(15.2)	(12.0)
MO	22.4	23.3	0.54	0.40	19.8	26.6	28.4	25.3
	(2.2)	(1.9)	(0.50)	(0.50)	(11.4)	(14.2)	(17.1)	(13.3)
MI	23.2	22.4	–	–	2.0	5.0	7.9	7.4
	(1.4)	(1.4)	–	–	(2.3)	(3.5)	(4.2)	(4.7)

Note July and August diurnal temperature range (DTR) are in °F. Jul90 and Aug90 are dummy variables indicating if the average maximum temperature is at least 90 °F for July and August, respectively. Rainfall events for May, June, July, and August are counts indicating at least one in/hr. of rainfall. Blank cells denote regressors that have been dropped due to insufficient variability. See text for data sources

Source NOAA GHCN data and NOAA COOP data

Fig. 1 Yield forecasts for Iowa and Illinois. Baseline specification, *t*-distributed yields. *Source* Authors' forecasts

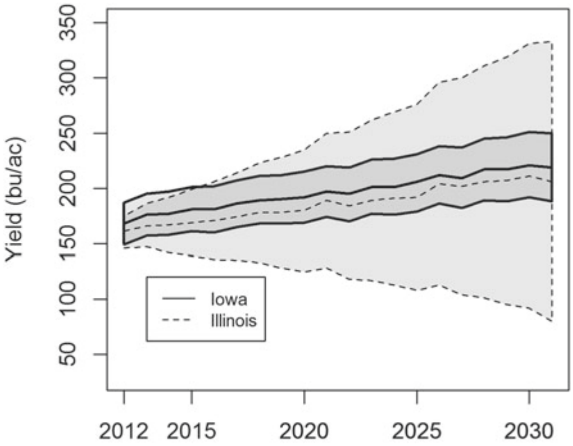


Fig. 2 Yield forecasts for Minnesota and Wisconsin. Baseline specification, *t*-distributed yields. *Source* Authors' forecasts

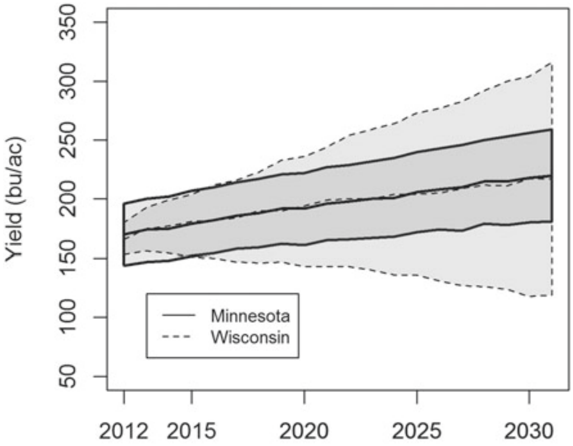
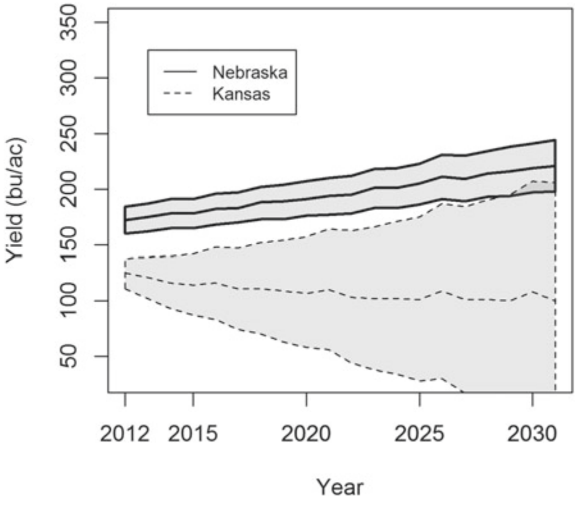


Fig. 3 Yield forecasts for Nebraska and Kansas. Baseline specification, *t*-distributed yields. *Source* Authors' forecasts



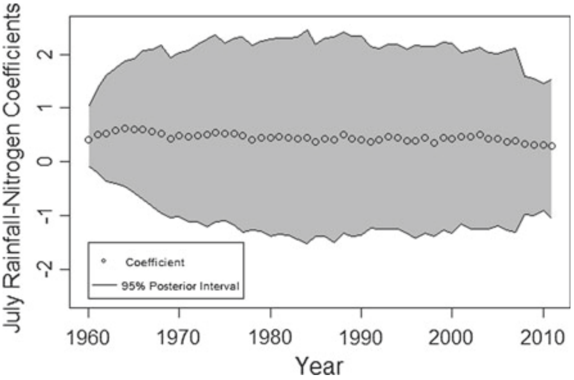


Fig. 4 July rainfall-nitrogen interaction coefficients. Iowa, baseline specification, beta-distributed yields. *Source* Authors' forecasts

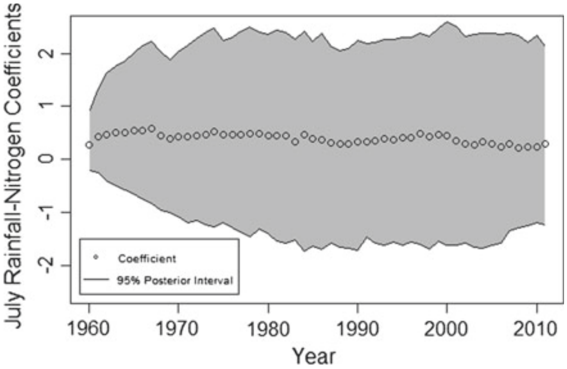


Fig. 5 July event-nitrogen interaction coefficients. Iowa, alternative specification, beta-distributed yields. *Source* Authors' forecasts

Table 3 Forecast mean absolute deviations

State	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	OLS, trend only	OLS, baseline	OLS, alternative	Student *t*-distribution, trend only	Student *t*-distribution, baseline	Student *t*-distribution, alternative	Beta distribution, trend only	Beta distribution, baseline	Beta distribution, alternative
IA	16.0	**12.0**	16.0	27.3	14.7	16.3	20.0	17.0	25.0
IL	33.3	36.3	33.0	**30.7**	33.7	31.3	33.0	33.0	31.3
NE	12.0	**11.3**	12.0	12.3	12.3	13.3	11.7	14.0	12.3
MN	10.0	**7.0**	7.7	19.3	13.0	10.7	13.7	11.0	8.3
IN	33.7	34.7	36.0	31.7	31.7	33.0	33.0	**31.3**	31.7
SD	16.3	17.3	**14.3**	15.7	19.3	**14.3**	15.7	16.0	20.3
WI	13.3	**11.0**	11.7	21.3	32.3	18.0	20.7	18.7	**11.0**
OH	22.3	20.0	26.7	19.3	14.0	25.0	20.0	**13.7**	27.3
KS	33.7	21.0	22.0	20.3	22.3	27.0	20.0	22.0	**19.3**
MO	37.3	**35.3**	41.7	37.7	36.3	37.7	37.0	36.3	39.3
MI	11.7	12.3	13.3	**8.7**	9.7	9.0	**8.7**	11.3	11.3

Note Mean absolute deviations (MAD) are calculated as the absolute value of the difference in forecasted and realized yields, averaged over 2012–2014. Units are bu/ac. For each state, the model with the lowest MAD is bolded. Column 1 uses OLS with only two regressors: intercept and linear time trend. Column 2 uses OLS with the baseline set of regressors. Column 3 uses OLS with the alternative set of regressors. Column 4 is the Dynamic Linear Model (DLM) with only two regressors: dynamic intercept and linear time trend. Column 5 is the DLM with the baseline set of regressors. Column 6 is the DLM with the alternative set of regressors. Column 7 is the beta-distributed model (using particle filtering) with only two regressors: dynamic intercept and linear time trend. Column 8 is the beta-distributed model with the baseline set of regressors. Column 9 is the beta-distributed model with the alternative set of regressors

Source Authors' calculations based on authors' estimates

References

1. Auffhammer, M., Hsiang, S.M., Schlenker, W., Sobel, A.: Using weather data and climate model output in economic analyses of climate change. Rev. Environ. Econ. Policy **7**(2), 181–198 (2013)
2. Barrows, G., Sexton, S., Zilberman, D.: Agricultural biotechnology: the promise and prospects of genetically modified crops. J. Econ. Perspect. **28**(1), 99–120 (2014)
3. Bureau of Reclamation (Reclamation): Downscaled CMIP3 and CMIP5 Climate and Hydrology Projections: Release of Downscaled CMIP5 Climate Projections, Comparison with Preceding Information, and Summary of User Needs. Technical memorandum. U.S. Department of the Interior (2013)
4. Claassen, R., Just, R.E.: Heterogeneity and distributional-form of farm-level yields. Am. J. Agric. Econ. **93**(1), 144–160 (2011)
5. Crosbie, T.M., Eathington, S.R., Johnson, Sr. G.R., Edwards, M., Reiter, R., Stark, S., Mohanty, R.G., Oyervides, M., Buehler, R.E., Walker, A.K., Dobert, R., Delannay, X., Pershing, J.C., Hall, M.A., Lamkey, K.R: Chapter 1. Plant breeding: past, present, and future. In: Lamkey, K.R., Lee, M. (eds.) Plant Breeding: The Arnel R. Hallauer International Symposium. pp. 3–50, Blackwell Publishing (2006)
6. da Silva, C., Migon, H.S., Correia, L.: Dynamic Bayesian beta models. Comput. Stat. Data Anal. **55**(6), 2074–2089 (2011)
7. Day, R.H.: Probability distributions of field crop yields. J. Farm Econ. **47**(3), 713–741 (1965)
8. de Gorter, H., Just, D.R.: The social costs and benefits of biofuels: the intersection of environmental, energy and agricultural policy. Appl. Econ. Perspect. Policy **32**(1), 4–32 (2010)
9. Durbin, J., Koopman, S.J.: Time Series Analysis by State Space Methods. Oxford University Press, Oxford (2012)
10. Environmental Protection Agency (EPA): Climate Change Indicators in the United States: Heavy Precipitation. http://www.epa.gov/climatechange/science/indicators/weather-climate/heavy-precip.html (2013)
11. Fernandez-Cornejo, J., Caswell M.: The first decade of genetically engineered Crops in the United States. Economic Information Bulletin No. 11. United States Department of Agriculture, Economic Research Service (2006)
12. Fernandez-Cornejo, J., Wechsler S.: Bt corn adoption by U.S. farmers increases yields and profits. Amber Waves. United States Department of Agriculture, Economic Research Service (2013)
13. Fernandez-Cornejo, J., Nehring, R., Osteen, C., Wechsler, S., Martin, A., Vialou, A.: Pesticide use in U.S. agriculture: 21 selected crops, 1960–2008. Economic Information Bulletin No. 124. United States Department of Agriculture, Economic Research Service (2014)
14. Fernandez-Cornejo, J., Wechsler, S., Livingston, M.: Adoption of genetically engineered crops by U.S. farmers has increased steadily for over 15 years. Amber Waves. United States Department of Agriculture, Economic Research Service (2014)
15. Foltz, J.D., Kim, K., Barham, B.: A dynamic analysis of university agricultural biotechnology patent production. Am. J. Agric. Econ. **85**(1), 187–197 (2003)
16. Groisman, P.Y., Knight, R.W., Karl, T.R.: Changes in intense precipitation over the Central United States. J. Hydrometeorol. **13**(1), 47–66 (2012)
17. Harri, A., Erdem, C., Coble, K.H., Knight, T.O.: Crop yield distributions: a reconciliation of previous research and statistical tests for normality. Appl. Econ. Perspect. Policy **31**(1), 163–182 (2009)
18. Heisey, P.W., Day-Rubinstein K.: Using crop genetic resources to help agriculture adapt to climate change: economics and policy. Economic Information Bulletin No. 139. United States Department of Agriculture, Economic Research Service (2015)
19. Hendricks, N.P., Peterson, J.M.: Fixed effects estimation of the intensive and extensive margins of irrigation water demand. J. Agric. Resour. Econ. **37**(1), 1–19 (2012)
20. Hennessy, D.A., Miranowski, J.A., Babcock, B.A.: Genetic information in agricultural productivity and product development. Am. J. Agric. Econ. **86**(1), 73–87 (2004)

21. Hornbeck, R., Keskin, P.: The historically evolving impact of the ogallala aquifer: agricultural adaptation to groundwater and drought. Am. Econ. J. Appl. Econ. **6**(1), 190–219 (2014)
22. Huffman, W.E., Shogren, J.F., Rousu, M., Tegene, A.: Consumer willingness to pay for genetically modified food labels in a market with diverse information: evidence from experimental auctions. J. Agric. Resour. Econ. **28**(3), 481–502 (2003)
23. Huffman, W.E.: Contributions of public and private R&D to biotechnology innovation. In: Carter, C.A., Moschini, G., Sheldon, I. (eds.) Genetically Modified Food and Global Welfare. Frontiers of Economics and Globalization, Vol. 10, pp. 115–147. Emerald Group Publishing Limited (2011)
24. Liu, J., West, M.: Combined parameter and state estimation in simulation-based filtering. In: Doucet, A., de Freitas, N., Gordon, N. (eds.) Sequential Monte Carlo Methods in Practice, pp. 197–224. Springer, Secaucus, NJ (2011)
25. Lobell, D.B., Hammer, G.L., McLean, G., Messina, C., Roberts, M.J., Schlenker, W.: The critical role of extreme heat for maize production in the United States. Nat. Clim. Chang. **3**, 497–501 (2013)
26. Lopes, H.F., Tsay, R.S.: Particle filters and Bayesian inference in financial econometrics. J. Forecast. **30**(1), 168–209 (2011)
27. Marshall, E., Aillery, M., Malcolm, S., Williams, R.: Climate change, water scarcity, and adaptation in the U.S. Fieldcrop Sector. Economic Research Report 201. United States Department of Agriculture, Economic Research Service (2015)
28. McFadden, J.R., Huffman, W.E.: Consumer valuation of information about food safety achieved using biotechnology: evidence from new potato products. Food Policy **69**, 82–96 (2017)
29. McFadden, J.R., Miranowski, J.A.: Climate change and U.S. corn yields: a dynamic Bayesian approach. In: Working paper. Department of Economics, Iowa State University (2014)
30. McFadden, J.R., Miranowski, J.A.: Climate change, technology, and U.S. corn yields. In: Working paper. Department of Economics, Iowa State University U.S. (2015)
31. McFadden, J.R., Miranowski, J.A.: Climate change impacts on the intensive and extensive margins of U.S. agricultural land. In: Working paper. Department of Economics, Iowa State University (2014)
32. McFadden, J.R., Miranowski, J.A.: Climate change: challenge and opportunity to maintain sustainable productivity growth and environment in a corn-soybean bioeconomy. AgBioForum **19**(2), 92–111 (2016)
33. McFadden, J.R.., Huffman, W.E.: Consumer demand for low-acrylamide-forming potato products: evidence from lab auctions. Am. J. Potato Res. (2017, forthcoming)
34. Migon, H.S., Gamerman, D., Lopes, H.F., Ferreira, M.A.: Chapter 19: dynamic models. In: Dey, D.K., Rao, C. (eds.) Bayesian Thinking: Modeling and Computation. Handbook of Statistics, Vol. 25, pp. 553–588. Elsevier, Amsterdam (2005)
35. Miranowski, J., Rosburg, A., Aukayanagul, J.: U.S. maize yield growth implications for ethanol and greenhouse gas emissions. AgBioForum **14**(3), 120–132 (2011)
36. Miranowski, J.A.: Technology forcing and associated costs and benefits of cellulosic ethanol. Choices **29**(1), 1–6 (2014)
37. National Agricultural Statistics Service (NASS): Quick Stats 2.0. http://www.quickstats.nass.usda.gov (2014)
38. National Oceanic and Atmospheric Administration (NOAA): Thunderstorms, Tornadoes, Lightning: Nature's Most Violent Storms. https://www.weather.gov/media/bis/TStorms_Tor_Lightning.pdf (2014)
39. Ortiz-Bobea, A., Just, R.E.: Modeling the structure of adaptation in climate change impact assessment. Am. J. Agric. Econ **95**(2), 244–251 (2013)
40. Pierce, D.W., Barnett, T.P., Santer, B.D., Gleckler, P.J.: Selecting global climate models for regional climate change studies. Proc. Nat. Acad. Sci. **106**(21), 8441–8446 (2009)
41. Pole, A., West, M., Harrison, J.: Applied Bayesian Forecasting and Time Series Analysis. Chapman & Hall, New York (1994)
42. Roberts, M.J., Schlenker, W., Eyer, J.: Agronomic weather measures in econometric models of crop yield with implications for climate change. Am. J. Agric. Econ. **95**(2), 236–243 (2013)

43. Romero-Lankao, P., Smith, J.B., Davidson, D.J., Diffenbaugh, N.S., Kinney, P.L., Kirshen, P., Kovacs, P., Villers Ruiz L.: North America. In: Barros, V.R., Field, C.B., Dokken, D.J., Mastrandrea, M.D., Mach, K.J., Bilir, T.E., Chatterjee, M., Ebi, K.L., Estrada, Y.O., Genova, R.C., Girma, B., Kissel, E.S., Levy, A.N., MacCracken, S., Mastrandrea, P.R., White, L.L. (eds.) Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 1439–1498. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2014)
44. Rosburg, A., McFadden, J., Miranowski, J.: Managing feedstock supply risk for the development of a U.S. stover biofuel industry. BioEnergy Res. (2017, forthcoming)
45. Schimmelpfennig, D., Lewandrowski, J., Reilly, J., Tsigas, M., Parry, I.: Agricultural adaptation to climate sustainability: issues of long run sustainability. Agricultural Economic Report No. 740. United States Department of Agriculture, Economic Research Service (1996)
46. Schlenker, W., Roberts, M.J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proc. Nat. Acad. Sci. **106**(37), 15594–15598 (2009)
47. Shi, G.: Bundling and licensing of genes in agricultural biotechnology. Am. J. Agric. Econ. **91**(1), 264–274 (2009)
48. Shi, G., Chavas, J., Stiegert, K.: An analysis of the pricing of traits in the U.S. corn seed market. Am. J. Agric. Econ. **92**(5), 1324–1338 (2010)
49. Smith, S., Cooper, M., Gogerty, J., Löffler, C., Borcherding, D., Wright K.: Chapter 6. Maize. In: Smith, S., Diers, B., Specht, J., Carver, B. (eds.) Yield Gains in Major U.S. Field Crops. pp. 125–171, American Society of Agronomy (2014)
50. United States Department of Agriculture (USDA). U.S. Department of Agriculture Climate Change Adaptation Plan. Government report (2014)
51. West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models. Springer, Secaucus (1997)
52. Xu, Z., Hennessy, D.A., Sardana, K., Moschini, G.: The realized yield effect of genetically engineered crops: U.S. maize and soybean. Crop. Sci. **53**(3), 735–745 (2013)
53. Zilberman, D., Liu, X., Roland-Holst, D., Sunding, D.: The economics of climate change in agriculture. Mitig. Adapt. Strateg. Glob. Chang. **9**(4), 365–382 (2004)
54. Zilberman, D., Zhao, J., Heiman, A.: Adoption versus adaptation, with emphasis on climate change. Annu. Rev. Resour. Econ. **4**, 27–53 (2012)

The Use of LCA for the Development of Bioenergy Pathways

Marcelle C. McManus

Abstract Bioenergy and biofuels are key to meeting renewable energy and carbon reduction targets. Life Cycle Assessment (LCA) techniques are being used, with varying success and consistency, to help determine the sustainability of the current fuels and pathways selected. In order to meet our longer term targets and pursue long term sustainability emerging processes and systems need to be examined, as well as existing processes. Designers recognise that a large percentage of impacts and costs are pre-ordained within the design stage; so it makes sense to use LCA at the start of the research process in order to minimise these. Determining impacts at this stage could also help select the most promising options with maximum sustainability/GHG reduction potential. At the same time policy makers are beginning to use LCA as a tool to help inform policy choices for future energy pathways. Never the less, there are various uncertainties involved with its use at early stage research level, and also the expansion of LCA to look at wider consequences of the use of a particular product or system. LCA is changing from a traditional, retrospective tool to a more dynamic, forward thinking tool. Whilst this brings a multitude of benefits in terms of ability to predict impacts and minimise these in advance, this method of LCA use is not without uncertainties and difficulties. This paper explores why LCA is important within the bioenergy context and highlights some of the benefits, disadvantages, and changes that are seen through its use.

Keywords Emerging LCA · Anticipatory · Bioenergy

1 Introduction

Many countries and regions have targets to increase the amount of bioenergy and biofuels in order to help minimise greenhouse gas emissions and meet climate change targets. In order to ensure that their use helps meet these targets it is important that

M.C. McManus (✉)
Department of Mechanical Engineering, University of Bath, England BA2 7AY, UK
e-mail: M.McManus@bath.ac.uk

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_17

their impact can be accurately, transparently and consistently measured. Life Cycle Assessment is an environmental management tool that is used to determine the impact towards a series of issues, such as climate change, resource use, acidification across a product or systems life; from production, and use, to disposal. It is increasingly used as a mechanism to help determine the sustainability of bioenergy systems and biofuel. The pathways from resource to fuel and use within bioenergy are many, and complex. The end users are focused on the availability of vehicle fuel, heat or electricity, but with bioenergy there are several methods available to produce these, see for example Fig. 1. Biomass resources vary from annual crops such as wheat, maize and sugar beet, to perennials such as miscanthus, switch grass, pine, spruce and residues and wastes, including forest residues, straw, municipal waste and waste oils. There are a similar number of conversion routes, including pyrolysis, gasification, esterification, digestion, etc., leading to a range of fuels such as biodiesel, bioethanol, bio-oil, bio-methane, methanol and hydrogen. LCA can be used to quantify the impact of these pathways with relatively high accuracy using attributional LCA. These impacts are commonly described in terms of energy and greenhouse balances, but other environmental impacts such as acidification, resource depletion and eutrophication can also be measured, and are often reported. Alongside the existing bioenergy pathways several more are under development. These can use novel feedstocks, such as algae, or new or rapidly developing conversion methods, such as the linocellulosic conversion to bioethanol. Many of these are at lab scale, meaning that LCAs are being performed at an earlier stage. This brings the associated benefits of being able to influence the process at an early process design stage, but with higher level of uncertainty due to the more experimental nature of the process. Despite, or perhaps because of, the increased uptake of bioenergy there has been a wide debate surrounding the sustainability of bioenergy, especially focusing around the food versus fuel debate (Royal Society [13]). For this reason second generation biofuels, which are made from biomass that doesnt directly compete with the food market (such as lignocellulosic bioethanol), are considered to be more beneficial.

2 Trends in Life Cycle Assessment

When LCA was initially developed in the 1970 to late 1990s it was a retrospective tool, predominantly used by industry in order to reduce resource use and waste production (Curran [3]; Hunt and Franklin [8]). The methodology was initially developed and published by the Society of Environmental Toxicology and Chemistry (SETAC) and these were then developed into a series of ISO standards in the 1990s. These standards were refined and amended in the 2000s. The method was developed to measure the impact of a product or system for which the data was currently or historically available (i.e. for a product in existence) for the purpose of decision making or reporting. Over the last years the way in which LCA is used has changed. This is primarily in two directions; wider towards a policy arena, and tightly focused around specific processes within early stage research. Over the last ten years a change in

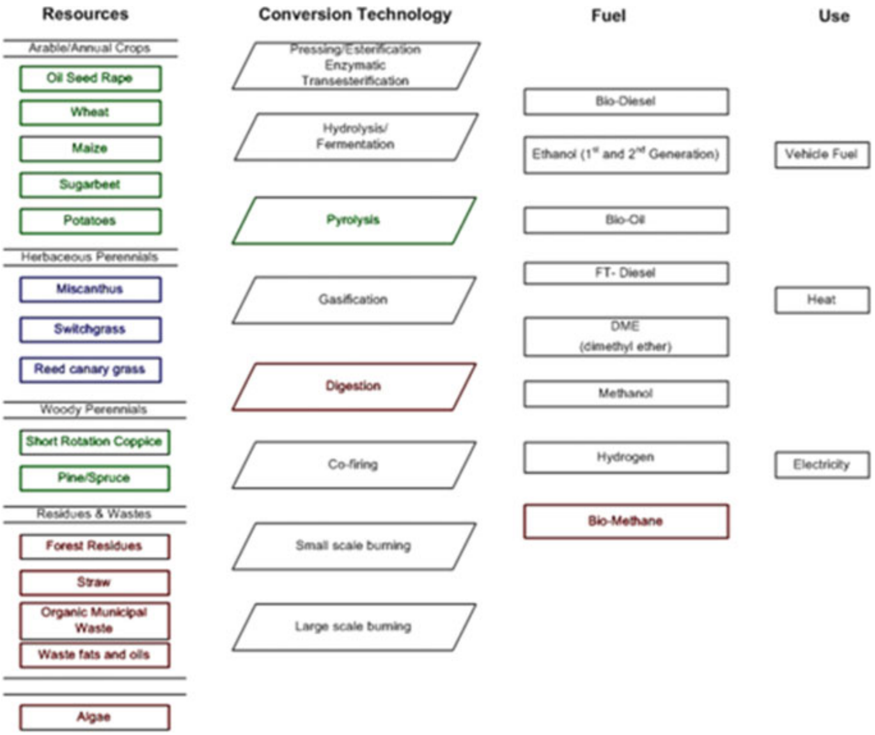


Fig. 1 Resources, Conversion technologies, fuels and uses for bioenergy

the way in which LCA has been described has also been seen. The more traditional type LCA is now often called attributional LCA (aLCA). That which looks wider, for example towards the implications of the use or expansion of a system, is called consequential LCA (cLCA). This move is reflected in the academic literature and the uptake of the tool (McManus and Taylor [11]). It is often presented in literature that 80% of all environmental effects associated with a product are determined during the design stage (Tischner et al. [19]), so the trend to increasingly use LCA at the early stages of research and design is relatively unsurprising (Fig. 2). Use at this stage enables the practitioner to explore options for minimising impact from the earliest stage of a product or systems life. LCA practitioners and researchers can work together to select the most environmentally benign materials and processes; hence reducing impact from the outset (e.g. Griffiths et al. [6]).

Whilst this can enable the reduction in negative impact, there are a number of methodological and practical difficulties that arise from using LCA in the determination of environmental load within the research stage of process development (Hetherington et al. [7]). One of the most significant issues when conducting early stage research based LCA is scalability. Lab based processes do not necessarily use the same processing stages as they would when commercialised, and efficien-

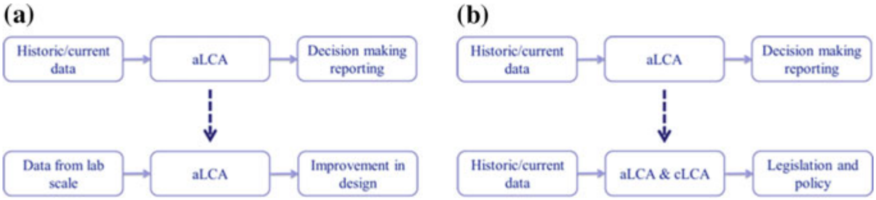


Fig. 2 Trends in LCA

cies are likely to be better at commercial scale. The resultant early stage LCA may have significantly more variables, complexities and scenarios than a “traditional” LCA (Hetherington, et al. [7]). Another issue commonly encountered in this type of research is the use of materials, enzymes etc. that have not previously been used; therefore the potential impact is particularly hard to predict for use within and LCA. In a field that is developing as rapidly as bioenergy this can be a significant issue. As an example, despite extensive research on both lab and small scale lignocellosic ethanol production, no large scale commercial lignocelluloses-to-ethanol facility has yet been brought into production. Therefore, technology uncertainty and potential commercial scale operation parameters also contribute to the knowledge gap when undertaking an LCA in this area (Spatari et al. [16]).

On the other side of the scale LCA is being used more outwardly, in a consequential approach, to help formulate policy. Consequential LCA is broader and explores the potential wider changes to the system that may arise from using the product in question (Sanchez et al. [14]). For example a consequential analysis of a biomass plant would examine the impacts of the production, use and disposal of the plant (and associated feestocks etc.), but could also include the impact of offsetting the energy that would have been alternatively used. As it takes into consideration a range of broader factors it is often used as a policy tool rather than a technology assessment (Plevin et al. [12]). It has been used most widely in the bioenergy arena (Taylor and McManus [18]).

As with the development of LCA into early stage research, the development of consequential LCA is not without problems. Many consequential LCAs have been developed from a number of attributional LCAs, with a number of smaller system studies being linked together to either add or offset each other, but some of these studies have been shown to produce misleading results (Bento and Klotz [2]). The systems that are under analysis, such as global biomass/land/energy systems, are complex; sometimes the only pragmatic option is to build the analysis from a series of smaller sub systems. Never the less, these dont necessarily reflect the complexity of the systems in question.

3 Developing Bioenergy Pathways: LCA Uncertainties

As bioenergy is promoted as a mechanism to provide low carbon energy it is clear that the impact of the bioenergy pathway selected is understood and that different options can be reliably compared. Bioenergy systems are complex. As is shown in Fig. 1 there are numerous feedstocks and conversion technologies. Many of the feedstocks have the potential to be grown for a multi purspose; i.e. after harvest they could be used for either bioenergy or another commodity, such as food, animal feed or the building industry. Such decisions will primarily be made on an economic basis, potentially bringing further uncertainties to any wide reaching LCA study in the area. The ISO standards oversee the general life cycle thinking approach to life cycle assessment, but there are also a number of tools that can be used and adopted to calculate the greenhouse gas emissions from numerous bioenergy systems, for example those developed to be used under the EU Renewable Energy Directive (RED) and the US based ones such as GREET and GHGenuis. Undertaking a full LCA requires expert knowledge, but the online tools can be used with a more cursory understanding of the underlying methodology. Whittaker et al [20] show that between the ISO standards, the GHG accounting methodologies such as PAS2015 and RED, and the online GHG tools there lies a significant decrease in consistency and transparency. This indicates the trade-off between the requirement of expert knowledge, and the use of quick GHG tools. Results from such tools (using the same inventory input data) range from just over 500 kg CO2 eq/ha to over 3000500 kg CO2 eq/ha. Some of the differences in approach result from differing allocation methods or the development of the counterfactual (what is displaced/not used) (Whittaker et al. [21]). Clearly, a consistent approach is required. A mechanism for understanding how the impacts from lab scale research is translated into impacts in commercial production; and a wider system for looking at global consequences is also required.

There is little research covering the implications in terms of consistency or predictability of moving between early and later product and system stages on environmental impact. Never the less, there are many disciplines from which LCA can learn. Business and technology development work in terms of technology readiness levels; from these funding and commercial predictions can be determined. It is certainly the case that as the technology matures there would be increased certainty of cost and impact, Fig. 3. However the manner of linkage is not yet established. Nor is it known whether there would or could be a repeatable mechanism for predicting impacts from lab scale research up to commercial scale research. More work is required in this area. Beyond the commercial processing impacts also lie the uncertainties associated with the use of the product in question. The development of LCA has been widely discussed over recent years. The expansion of LCA from an attributional approach only, to those that look at the wider consequences (cLCA) and all the studies, tools and methods that lie between the two have been widely discussed (see for example Whittaker et al. ([20],[21]). While there are problems and issues with the simplification of any system into a model, the use of such models is perhaps the only way in which we can determine likely impacts of our activities. Although aLCA and

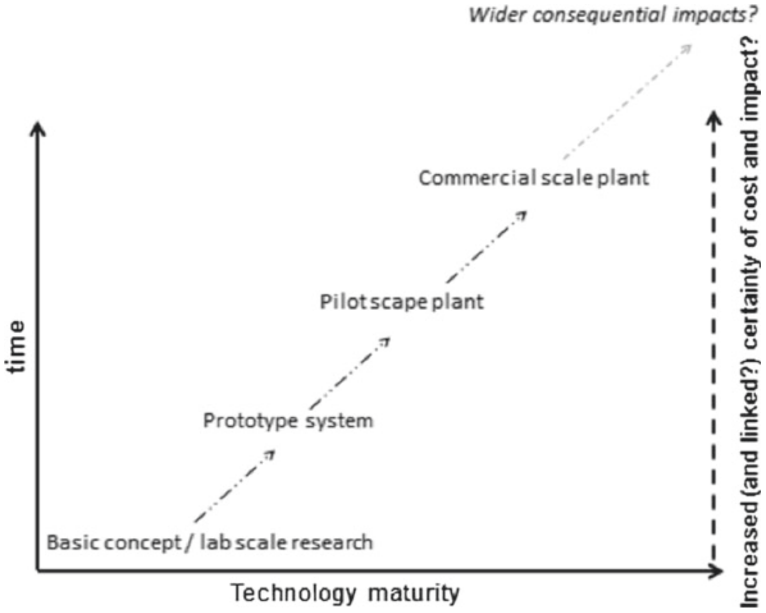

Wider consequential impacts?
Commercial scale plant
Pilot scape plant
Prototype system
Basic concept / lab scale research
time
Technology maturity
Increased (and linked?) certainty of cost and impact?

Fig. 3 Technology maturity and potential certainty in LCA impacts

cLCA have been criticised for lack of consistency and transparency at times (see for example McManus et al. and Plevin et al. [12]) there really is currently no better way to model the complexities associated with the production and use of, for example, biofuel. Recognition of the weaknesses of the current system and tools does exist, and the opportunity that currently exists to improve is crucial. As the systems expand in the more consequential LCAs the adaptation of knowledge from other disciplines is required and how we need to examine how a model, or combination of models, can used in order to answer complex and dynamic questions whilst recognising both strengths and weaknesses of the modelling frameworks and available data (Suh and Yang [17]).

As bioenergy markets expand it is likely that the global systems will maintain complexity that is difficult to model. It is also likely that the research into novel ways of extracting energy from biomass will continue at a rapid pace. Expanding on Figs. 2 and 4 explores the option of moving from a tight attributional type LCA to a wider consequential one based primarily on the type of data coming from lab scale research. This is beginning to be seen in the public discourse surrounding bioenergy as speculation of future scenarios and how new types of biofuels and bioenergy might help our future demand increases. Strategic policy making that encompasses thoughts of potential impacts is to be highly commended. However, it is clear that LCA is at a point where lessons from other sectors could possibly be incorporated and that a clear indication of the level of speculation and uncertainty associated with any such study should be highlighted.

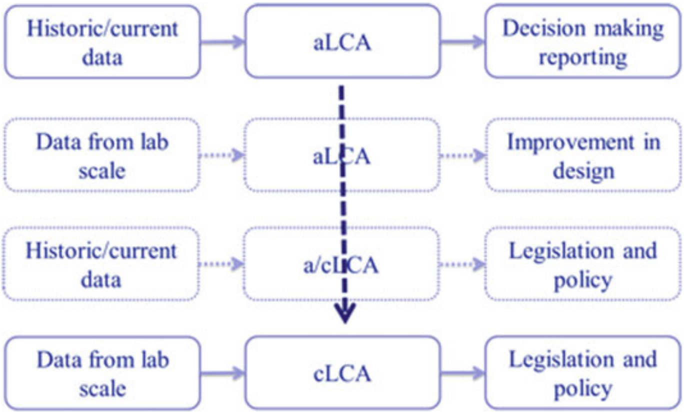


Fig. 4 Current and potential trends in LCA development

4 Conclusions

Life Cycle Assessment is a tool that can be extremely helpful to determine impacts across a wide sector. It is of particular use in the bioenergy area; where many policies and legislative mechanisms are developing that require the use of LCA. There is a requirement that any replacement for our current fossil fuel system has a beneficial comparative impact in terms of greenhouse gas emissions. LCA is an excellent tool to measure this. The use of LCA is changing rapidly; and at this time of transition there is an opportunity to reflect and learn from other sectors. No model can accurately map the complexities of potential positive or negative impacts associated with the production and use of bioenergy. Life Cycle Assessment is the closest we have to a tool that can predict impacts and enable us to minimise and reduce them. It has been a very successful tool, with use in policy making and legislation increasing exponentially over a short period of time. Never the less, the way in which it is used is being stretched. Recognition of this may mean that the fragility of the model can be overcome and LCA will emerge a stronger and ever increasingly used tool. But if these issues and problems are ignored it is possible that the tool will become increasingly mis-used and the results mis-interpreted with regrettable effects on both bioenergy and the LCA tool itself.

Acknowledgements The author would like to thank her funders; the BBSRC UK/US partnership grant BB/J020184/1, the EPSRC Supergen Bioenergy Hub (EP/J017302/1), BBSRC LACE programme (BB/G01616X/1), and the University of Bath. This work was originally presented at the Berkeley Bioeconomy Conference in March 2014 and has been augmented and updated for this publication.

References

1. ANL 2014. GREET. The greenhouse gases, regulated emissions, and energy use in transportation (GREET) model. https://greet.es.anl.gov/main
2. Bento, A.M., Klotz, R.: Climate Policy Decisions Require Policy-based Lifecycle Analysis. Environmental Science and Technology, ACS (2014)
3. Curran, M.A.: Broad-based environmental life cycle assessment. Environ. Sci. Technol. **27**(3), 430–436 (1993)
4. EU: Renewable Energy Directive. Directive 2009/28/EC of the European Parliament and of the Council of 23 April on the Promotion of the Use of Energy from Renewable Sources and Amending and Subsequently Repealing Directives 2001/77/EC and 2003/30/EC (2009)
5. GHGenius. http://www.ghgenius.ca/downloads.php (2014). Accessed 12 Dec 2014
6. Griffiths, O.G., Owen, R.E., OByrne, J.P. Mattia, D., Jones, M.D., McManus M.C.: LCA Using Life Cycle Assessment to Measure the Environmental Performance of Catalysts and Directing Research in the Conversion of CO2 into Commodity Chemicals: a look at the potential for fuels from thin-air RSC Advances, vol. 30 (2013)
7. Hetherington, A., Borrion, A.L., Griffiths, O.G., McManus, M.C.: The use and implications of LCA in early stage research. J. Life Cycle Assess. **19**(1), 130–143 (2014)
8. Hunt, R.G., Franklin, W.: LCAHow it came about. Int. J. Life Cycle Assess. **1**(1), 4–7 (1996)
9. ISO. Environmental management life cycle assessment principles and framework. International Standards Organization, Second Edition, EN ISO 14040 (2006)
10. ISO. Environmental management life cycle assessment requirements and guidelines, International Standards Organization, EN ISO 14044 (2006)
11. McManus and Taylor (submitted) The Changing Nature of Life Cycle Assessment. Submitted to Biomass and Bioenergy
12. Plevin, R.J., Delucchi, M.A., Creutzig, F.: Using attributional life cycle assessment to estimate climate change mitigation benefits misleads policy makers. J. Ind. Ecol. **18**(1), 73–83 (2014)
13. Royal Society Sustainable Biofuels Prospects and Challenges. Policy Document 01/08. ISBN 9780854036622. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2008/7980.pdf (2008)
14. Sanchez, S.,Woods, J., Akhurst, M., Brander, M., O'Hare, M., Dawson, T.P., Edwards, R., Liska, A.J., Malpas, R.: Accounting for indirect land-use change in the life cycle assessment of biofuel supply chains J. R. Soc. Interface 7 June 2012 vol. 9 no. 71 1105-1119. http://rsif.royalsocietypublishing.org/content/9/71/1105.short (2012)
15. Society of Environmental Toxicology and Chemistry (SETAC) Guidelines for Life-Cycle Assessment: A code of practice. Based on a workshop at Sesimbra, Portugal, March 31 April 3. Brussels and Pensacola, Florida, USA (1993)
16. Spatari, S., Bagley, D.M., MacLean, H.L.: Life cycle evaluation of emerging lignocellulosic ethanol conversion technologies. Bioresour. Technol. **101**, 654–667 (2010)
17. Suh, S., Yang, Y.: On the uncanny capabilities of consequential LCA. Int. J. Life Cycle Assess. **19**, 1179–1184 (2014). doi:10.1007/s11367-014-0739-9
18. Taylor, C.M., McManus, M.C.: The Evolving Role of LCA in Bioenergy Policy, Bioenergy Connection 2013 VOL. 2.3 MAGAZINE. Energy BioSciences Institute (2013). http://energybiosciencesinstitute.org/publications/2013-vol-23-magazine#overlay-context=
19. Tischner, U., Schmincke, E., Rubik, F., Prosler, M.: How to Do Ecodesign?: A Guide for Environmentally and Economically Sound Design, Edited by the German Federal Environmental Agency, Verlag form (Praxis) ISBN: 3898020258/9783898020251 (2000)
20. Whittaker, C., McManus, M.C., Smith, P.: A comparison of carbon accounting tools for bioenergy and for whole farms. Environ. Model. Softw. (2013)
21. Whittaker, C., McManus, M.C., Hammond, G.P.: Greenhouse gas reporting for biofuels: a comparison between the RED, RTFO and PAS2050 methodologies. Energy Policy (2011). doi:10.1016/j.enpol.2011.06.054

Optimal Control of Stochastic Hybrid Models in the Framework of Regime Switches

E. Savku, N. Azevedo and G.W. Weber

Abstract Stochastic Hybrid Systems with Jumps (SHSJs) are natural, powerful and efficient candidates to model abrupt changes in financial markets as a consequence of their heterogenous nature, especially, in the framework of regime switches. The internet bubble and the 2008–2009 economic crash forced researchers and practitioners to develop new portfolio models rather than relying on the traditional ones, and it is seen that regime-switching asset allocation significantly improves the performance compared with unconditional static alternatives. In addition to regime switches, which may occur in economics and finance, our study is also applicable for sudden paradigm changes. This means that cultural and also societal transformations, as far as they can be represented by stochastic dynamics, may undergo "switches" as investigated in this chapter. Let us mention that, in the real world, cultural, societal, economical and financial developments are closely related with each other. In this respect, the present chapter opens a new and wider view on the world of today and tomorrow. In this chapter, we start by introducing SHSJs and providing an extension for Bellman's optimality principle before moving on to a Markov-switching jump diffusion stochastic differential equation in a finite time horizon. Additionally, the corresponding Hamilton-Jacobi-Bellman equation is given and a consumption-investment strategy for a jump-diffusion financial market consisting of one risky and one risk-free asset whose coefficients are assumed to depend on the state of a continuous-time Markov process is presented. A more general model for portfolio optimization with a time delay structure and some numerical approaches for such problems are discussed as an outlook on SHSJs with delay.

E. Savku (✉) · G.W. Weber
Institute of Applied Mathematics, METU, 06800 Ankara, Turkey
e-mail: esavku@metu.edu.tr

G.W. Weber
e-mail: gweber@metu.edu.tr

N. Azevedo
CEMAPRE, ISEG, Universidade de Lisboa, ESEIG, Polytechnic Institute of Porto, Lisbon, Portugal
e-mail: nazevedo@iseg.utl.pt

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_18

Keywords Stochastic hybrid systems · Markov regime-switches · Jump processes · Dynamic programming principle · Hamilton-Jacobi-Bellman equations · Application to finance

1 Introduction

A Stochastic Hybrid System (SHS) is a continuous-time dynamic with discrete components. Throughout this chapter we will assume that the continuous variable x belongs to a subset of an Euclidean space, $\mathbf{X} \subseteq \mathbb{R}^d$ and discrete variable q belongs to a countable set Θ. The main difference between these two variables is the way that they evolve. The continuous variable evolves according to a stochastic differential equation (SDE) or a stochastic differential equation with jumps (in the case of jump diffusions). In the latter case such a system is called Stochastic Hybrid System with Jumps (SHSJ). Hence, unlike conventional hybrid systems, stochastic uncertainty arises as represented by diffusion term in SHS and by both diffusion and jump parts in the case of SHSJs. The discrete dynamics produce transitions in both continuous and discrete state variables, which are called *hybrid states* altogether. The switching between two discrete states is governed by a probability law or occurs when the continuous state hits the boundary of its state space. Transitions, which occur when the continuous state variable hits the boundary of the state space are called *forced transitions* and those which occur probabilistically according to a state dependent rate are called *spontaneous transitions* [1]. Whenever a switching occurs, the hybrid state is reset instantly to a new state according to a probability law which may depend itself on the past hybrid state. Thus a SHS (and a SHSJ) can be considered as an interleaving between a finite or countable family of diffusion processes (and jump processes) and a jump process [8, 22].

Deterministic and non-deterministic hybrid systems have been involved in a wide range of behaviors encountered in practice but their limitations make them too coarse for certain applications. In particular, while deterministic hybrid systems do not allow uncertainty in the model, non-deterministic hybrid systems provide no way of distinguishing between solutions which implies only worst case analysis, a kind of "Yes-No" question, is possible. For example, a safety question of an air traffic management problem for non-deterministic hybrid systems admits only one of two answers: The system is safe (if none of the solutions of the system ever reaches an unsafe state - the "Yes" case), or the system is not safe (if at least one solution reaches some unsafe state - the "No" case).

On the other hand, the inclusion of randomness makes the analysis of SHS (and SHSJ) more challenging and complicated, it allows for applications in a diverse range of scientific and engineering fields. Furthermore, when their heterogeneous structure is taken into account as well, a SHS (and a SHSJs) is a powerful tool to model dynamical systems with multiple modes, e.g., air traffic management [10, 34], biological networks [2], automotive systems [4], manufacturing systems and fault tolerant control [20, 21].

Moreover, by combining their efficiency to capture abrupt changes with the highly used stochastic optimal control theory, SHS (and SHSJ) arise in numerous applications not only in finance and economy, but also in actuary science (e.g., insurance pricing [12]). When we consider the interaction between the discrete components and continuous dynamics, it is clearly seen that such systems are useful and appropriate to represent the shifts in the financial market behaviors such as, a shift from a "bull" day to a "bear" day or a shift from a "volatile" day to a "choppy" trading day and also the movements of the market between various states, e.g., growth, crisis, range bounds or bubbles. In particular, if we focus on regime switches, there are several works which take into consideration regime switches, such as option pricing and risk minimization [13–15], consumption-investment problems [3], determining optimal selling rules [35] and optimal asset allocation [36].

This book chapter is organized as follows. In Sect. 2, SHSs are defined and their basic properties are provided. In Sect. 3, SHSJs are described as solutions of SDEs. In Sect. 4, dynamic programming techniques for a Markov-switching jump-diffusion are introduced and an application in finance is given in the framework of a consumption-investment problem. Especially, in Sects. 3 and 4, we present the results obtained by Azevedo et al. [3]. Section 5 is devoted to present some conclusions and an outlook to future work.

The authors would like to thank Prof Dr Diogo Pinheiro for his valuable comments and useful discussions.

2 General Stochastic Hybrid Model

In this section, firstly we will provide the general definition of Stochastic Hybrid Systems.

Definition 1 A *Stochastic Hybrid System* (SHS) is a collection $H = (Q, X, lnv, f, g, G, R)$, where

1. θ is a discrete variable taking values in a countable set $\Theta = \{q_1, q_2, ...\}$.
2. X is a continuous variable taking values in $\mathbf{X} = \mathbb{R}^N$ for some $N \in \mathbb{N}$.
3. $lnv : \Theta \to 2^{\mathbf{X}}$ assigns to each $q \in \Theta$ an invariant open subset of $\mathbf{X}$.
4. $f, g : \Theta \times \mathbf{X} \to T\mathbf{X}$ are vector fields.
5. Let $E = \Theta \times \Theta$ and denote by $G : E \to 2^{\mathbf{X}}$ that the map assigns to each $e \in E$ a guard $G(e)$ such that:
   - For each $e = (q, q') \in E$, $G(e)$ is a measurable subset of $\partial lnv(q)$,
   - For each $q \in \Theta$, the family $(G(e) : e = (q, q')$ for some $q' \in \Theta)$ is a disjoint partition of $\partial lnv(q)$.
6. $R : E \times \mathbf{X} \to P(\mathbf{X})$ assigns to each $e = (q, q') \in E$ and $x \in G(e)$ a reset probability kernel on $\mathbf{X}$ concentrated on $lnv(q')$.

Here, $2^{\mathbf{X}}$ denotes all subsets of $\mathbf{X}$ and $P(\mathbf{X})$ represents the family of all probability measures on $\mathbf{X}$. Moreover, for any measurable set $A \subset lnv(q')$, $R(e, x)(A)$ is a measurable function in x. The measurability assumption on R ensures that the events are measurable with respect to the underlying σ-field, thus their probabilities are well-defined.

Let us present another fundamental definition.

Definition 2 (*Stochastic Execution*) A stochastic process $(X(t), \theta(t))_{t\geq 0}$ in $\mathbf{X} \times \Theta$ is called a *stochastic execution* if and only if there exists a sequence of stopping times $\tau_0 = 0 \leq \tau_1 \leq \tau_2 \leq \ldots$ such that for each $n \in \mathbb{N}_0$:

1. In each interval $[\tau_n, \tau_{n+1})$, $\theta(t) \equiv \theta(\tau_n)$ is constant and $X(t)$ is a continuous solution to the SDE:

$$dX(t) = f(X(t), \theta(\tau_n))dt + g(X(t), \theta(\tau_n))dW(t),$$

   where $(W(t))_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}$.
2. $\tau_{n+1} = \inf\{t \geq \ \tau_n : X(t) \notin lnv(\theta(\tau_n))\}$.
3. $X(\tau_{n+1}^-) \in G(\theta(\tau_n), \theta(\tau_{n+1}))$, where $X(\tau_{n+1}^-)$ denotes the $\lim_{t\uparrow\tau_{n+1}} X(t)$.
4. The probability distribution of $X(\tau_{n+1})$ given $X(\tau_{n+1}^-)$ is governed by the law $R(e_n, X(\tau_{n+1}^-))$, where $e_n = (\theta(\tau_n), \theta(\tau_{n+1})) \in E$.

We now give the definition of an embedded Markov process which provides a way to get an explicit expression of the stochastic execution for a stochastic hybrid system if the problem is a reachability analysis of the discrete transitions [23].

Definition 3 (*Embedded Markov Processes*) Let τ_n be the sequence of stopping times introduced in Definition 2, $\theta_n \overset{\triangle}{=} \theta(\tau_n)$ and $X_n \overset{\triangle}{=} X(\tau_n)$. Then $((X_n : \theta_n) : n \geq 0)$ is called an *embedded Markov process* for the stochastic execution $(X(t), \theta(t))_{t\geq 0}$.

Based on these definitions, one can see that a typical stochastic execution starts from (X_0, θ_0) and the continuous state evolves according to the SDE;

$$dX(t) = f(X(t), \theta_0)dt + g(X(t), \theta_0)dW(t), \qquad X(0) = X_0,$$

up to time τ_1 when $X(t)$ first hits $\partial lnv(\theta_0)$. Then, depending on the hitting position $X(\tau_1^-)$, the discrete state jumps to $\theta(\tau_1) = \theta_1$ and the continuous state is reset randomly to $X(\tau_1) = X_1$ according to the probability distribution $R(e, X(\tau_1^-))(\cdot)$, and the same process is repeated with (X_1, θ_1) replacing (X_0, θ_0), and so on.

When we consider a SHSJ, a jump diffusion will be encountered instead of the SDE in the above definitions and the process $X(t)_{t\geq 0}$ jumps from one jump-diffusion path to another as the discrete component $\theta(t)_{t\geq 0}$ switches from one discrete state to another.

3 Stochastic Hybrid Model with Jumps as Solution of SDEs

Ghosh and Bagchi [22] proposed two hybrid models, both of which permit diffusion and hybrid jump. In the first model, they constructed a Markov process $(X, \theta) =$

$(X(t), \theta(t))_{t\geq 0}$, where X is governed by a SDE of Itô-Skorohod type with the drift coefficient, the diffusion matrix and the jump function depending on the discrete component θ. In the second model, they provided a study of a more general SHS where instantaneous jumps occur when X hits the specified subsets of its evolution sets and the destination of X is determined by a pre-determined map. Furthermore, they provide the conditions to guarantee that there exists unique strong solution of the SDE related to each model. Krystul and Blom [25] presented two other models and provided a hierarchy among the hybrid models in [8, 9, 22, 25].
Let us introduce SHSJs as a solution of SDE with jumps modulated by an exogenous process θ.

Let $(X, \theta) = (X(t), \theta(t))_{t\geq 0}$ be a switching jump diffusion taking values in $\mathbb{R}^n \times \Theta$, where $\Theta = \{e_1, e_2, ..., e_N\}$ is a finite set. Here, $e_i \in \mathbb{R}^N$ represents the ith unit vector for each $i = 1, 2, ..., N$. We define (X, θ) by the following SDE of Itô-Skorohod type:

$$dX(t) = f(X(t), \theta(t))dt + g(X(t), \theta(t))dW(t) + \int_{\mathbb{R}^d} h_1(X(t_-), \theta(t_-), z)q_1(dt, dz) \tag{1}$$

$$+ \int_{\mathbb{R}^d} h_2(X(t_-), \theta(t_-), z)p_2(dt, dz),$$

$$d\theta(t) = \int_{\mathbb{R}^d} c(X(t_-), \theta(t_-), z)p_2(dt, dz). \tag{2}$$

Let assume,

1. $X(0)$ be an $\mathbb{R}^n$-valued random variable.
2. $\theta(0)$ be a Θ-valued random variable.
3. $W = (W(t) : t \in [0, \infty))$ be an m-dimensional Brownian motion.
4. $q_1(dt, dz)$ be a martingale random measure associated to a Poisson random measure p_1 with intensity $dt \times \nu_1(dz)$.
5. $p_2(dt, dz)$ be a Poisson random measure with intensity $dt \times \nu_2(dz) = dt \times dz_1 \times \overline{\mu}(d\underline{z})$ where $\overline{\mu}$ is a probability measure on $\mathbb{R}^{d-1}$, $z_1 \in \mathbb{R}$ and $\underline{z} \in \mathbb{R}^{d-1}$ refers to all components except the first one of $z \in \mathbb{R}^d$.

Let us introduce the coefficient functions as follows:

$$f : \mathbb{R}^n \times \Theta \to \mathbb{R}^n, \quad g : \mathbb{R}^n \times \Theta \to \mathbb{R}^{n\times m}, \quad h_1 : \mathbb{R}^n \times \Theta \times \mathbb{R}^d \to \mathbb{R}^n,$$
$$h_2 : \mathbb{R}^n \times \Theta \times \mathbb{R}^d \to \mathbb{R}^n, \quad \phi : \mathbb{R}^n \times \Theta \times \Theta \times \mathbb{R}^{d-1} \to \mathbb{R}^n,$$
$$\lambda : \mathbb{R}^n \times \Theta \times \Theta \to \mathbb{R}_+, \quad c : \mathbb{R}^n \times \Theta \times \mathbb{R}^d \to \mathbb{R}^N.$$

Furthermore, we define the measurable mappings $\sum_k : \mathbb{R}^n \times \Theta \to \mathbb{R}_+$ for all $k = 1, 2, .., N$, in the following way:

$$\sum_k(x, e_i) = \begin{cases} \sum_{j=1}^k \lambda(x, e_i, e_j), & \text{if } k > 0 \\ 0, & \text{if } k = 0 \end{cases}, \tag{3}$$

$$c(x, e_i, z) = \begin{cases} e_i - e_j, & \text{if } z_1 \in (\sum_{j-1}(x, e_i), \sum_j (x, e_i)] \\ 0, & \text{otherwise} \end{cases} , \tag{4}$$

and

$$h_2(x, e_i, z) = \begin{cases} \phi(x, e_i, e_j, \underline{z}), & \text{if } z_1 \in (\sum_{j-1}(x, e_i), \sum_j (x, e_i)] \\ 0, & \text{otherwise} \end{cases} , \tag{5}$$

referring to some suitable j. The jump size of X and the new value of θ at the impulse times generated by the Poisson random measure p_2 are determined by the functions in Eqs. (5) and (4) respectively. There may occur three different situations: Simultaneous jump of X and θ (hybrid jump case), switch of θ only and jump of X only [25].

Definition 4 (*Hybrid Jump*) A *Hybrid Jump* is a continuous-valued jump that happens simultaneously with a mode switch such that its size depends on the mode value prior and after the switch.

Let us present the technical assumptions on the coefficients of the system of SDEs in Eqs. (1) and (2).

1. There exists a constant l such that for all $i = 1, 2, ..., N$, we have that

$$|f(x, e_i)|^2 + |g(x, e_i)|^2 + \int_{\mathbb{R}^d} |h_1(x, e_i, z)|^2 \, \nu_1(dz) \le l(1 + |x|^2).$$

2. For any $r > 0$ one can specify a constant l_r such that for $|x| \le r$, $|y| \le r$ and $i = 1, 2, ..., N$, the following condition holds:

$$\begin{aligned} &|f(x, e_i) - f(y, e_i)|^2 + |g(x, e_i) - g(y, e_i)|^2 \\ &+ \int_{\mathbb{R}^d} |h_1(x, e_i, z) - h_1(y, e_i, z)|^2 \, \nu_1(dz) \le l_r \, |x - y|^2 . \end{aligned}$$

3. Function c satisfies Eqs. (3), (4) and for all $t > 0$, $i, j = 1, 2, ..., N$, $\lambda(e_i, e_j, \cdot)$ are bounded and measurable, and $\lambda(e_i, e_j, \cdot) \ge 0$.
4. h_2 satisfies Eqs. (3) and (5), and for all $t > 0$, $i, j = 1, 2, ..., N$:

$$\int_0^t \int_{\mathbb{R}^d} \left|\phi(x, e_i, e_j, \underline{z})\right| p_2(ds, dz) < \infty, \; P - a.s.$$

Krystul and Blom (Theorem 4.1 in [25]) proved that the system of SDEs in Eqs. (1) and (2) has a unique strong solution. Let us express this by the following theorem.

Theorem 1 *Let us assume that the above technical assumptions hold and p_1, p_2, W, $X(0)$ and $\theta(0)$ be independent. Then the system of SDEs in Eqs. (1) and (2) has a unique strong solution which is a semimartingale.*

4 Dynamic Programing for a Markov-Switching Jump-Diffusion

The dynamic programming technique was firstly introduced by Richard Bellman in the 1950s to deal with calculus of variations and optimal control problems [5, 6]. A very complete treatment of the modern theory of optimal control problems can be found in the monographs by Fleming and Soner [18] and Øksendal and Sulem [30]. Furthermore, stochastic optimal control theory has been successfully applied to financial mathematics. Framstad et al. [19] gave a maximum principle for the optimal control of jump diffusions, showed its connections to dynamic programming and also provided applications to financial optimization problems. On the other hand, Bensoussan and Menadi [7] presented dynamic programming techniques for SHS, a more general class of stochastic processes.

Azevedo et al. [3] proved an extension of Bellman's optimality principle for a large class of stochastic optimal control problems whose state variable dynamics are given by a Markov-switching jump diffusion stochastic differential equation.

4.1 Problem Formulation

Let $T > 0$ be a deterministic finite horizon and let $(\Omega, \mathscr{F}, \mathbb{F}, P)$ be a complete filtered probability space with filtration $\mathbb{F} = (\mathscr{F}_t : t \in [0, T])$ satisfying the usual conditions, i.e., $\mathbb{F}$ is an increasing, right-continuous filtration and $\mathscr{F}_0$ contains all P-null sets. For each $d \in \mathbb{N}$, let $\mathbb{R}_0^d = \mathbb{R}^d \setminus \{0\}$ and let $\mathscr{B}_0^d$ be the Borel σ-field generated by the open subsets O of $\mathbb{R}_0^d$ whose closure does not contain 0.

Let $(W(t) : t \in [0, T])$ be an M-dimensional Brownian motion defined on $(\Omega, \mathscr{F}, \mathbb{F}, P)$ over $[0, T]$ and adapted to the filtration $\mathbb{F}$ and let $(\theta(t) : t \in [0, T])$ be a continuous-time Markov process with a finite space state $\Theta = \{a_1, a_2, ..., a_n\}$ and $Q = (q_{ij})_{i,j\in\Theta}$ be the generator. Let define the counting process N_{ij} as

$$N_{ij}(t) = \sum_{0<s\leq t} I_{\{\theta(s_-)=i\}} I_{\{\theta(s)=j\}} \qquad (t \geq 0),$$

where I_A denotes the indicator function of a set A. Here, $N_{ij}(t)$ gives the number of jumps of the Markov process θ from state i to j up to time t. Let us define the intensity process by

$$\lambda_{ij}(t) = q_{ij} I_{\{\theta(t_-)=i\}}$$

and the martingale process by

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \lambda_{ij}(s)ds.$$

Furthermore, let us define a K-dimensional Lévy process $(\eta(t) : t \in [0, T])$ with Poisson random measure $J(t, A)$ whose intensity (Lévy measure) is $\nu(A) = E[J(1, A)]$. Hence, the compensated Poisson random measure of $\eta(t)$ is defined by

$$\tilde{J}(t, A) = J(t, A) - t\nu(A).$$

For each $s \in [0, T)$, we denote the set of 8-tuples $(\Omega, \mathscr{F}, \mathbb{F}, P, W(\cdot), \theta(\cdot), \eta(\cdot), u(\cdot))$ by $U^w[s, T]$ which is called the set of *weak admissible controls*. The following conditions are assumed to hold for $U^w[s, T]$:

1. $(\Omega, \mathscr{F}, \mathbb{P})$ is a complete probability space.
2. $\mathbb{F} = (\mathscr{F}_t^s)_{t \geq s}$ is a right-continuous filtration.
3. $(W(t) : t \in [s, T])$ is an M-dimensional standard Brownian motion defined on $(\Omega, \mathscr{F}, P)$ over $[s, T]$ and adapted to the filtration $\mathbb{F}$.
4. $(\theta(t) : t \in [s, T])$ is a continuous-time Markov process with a finite state space Θ and adapted to the filtration $\mathbb{F}$.
5. $(\eta(t) : t \in [s, T])$ is a K-dimensional Lévy process defined on $(\Omega, \mathscr{F}, P)$ over $[s, T]$ and adapted to the filtration $\mathbb{F}$.
6. $u : [s, T] \times \Omega \to U$ is an $(\mathscr{F}_t^s)_{t \geq s}$-adapted process on $(\Omega, \mathscr{F}, P)$.
7. Under $u(\cdot)$, for any $y \in \mathbb{R}^N$ and $i \in \Theta$, the SDE in Eq. (6) admits unique solution $X(\cdot)$ on $(\Omega, \mathscr{F}, \mathbb{F}, P)$.

Furthermore, for any $(s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta$ and $u(\cdot) \in U^w[s, T]$, we consider the SHSJ in the following form:

$$\begin{aligned} dX(t) &= f(t, X(t_-), \theta(t_-), u(t_-))dt + g(t, X(t_-), \theta(t_-), u(t_-))dW(t) \\ &\quad + \int_{\mathbb{R}_0^K} h(t, X(t_-), \theta(t_-, u(t_-), z)\tilde{J}(dt, dz) \qquad (t \in [s, T]), \\ X(s) &= y, \qquad \theta(s) = i. \end{aligned} \tag{6}$$

Our objective (performance) functional is given by the following expected utility:

$$J(s, y, i; u(\cdot)) = E\left[\int_s^T L(t, X_{s,y,i}(t; u(\cdot)), \theta_{s,i}(t), u(t))dt + \Psi(T, X_{s,y,i}(T; u(\cdot)), \theta_{s,i}(T))\right] \tag{7}$$

where $X_{s,y,i}(t, u(\cdot), \theta_{s,i}(t)) \in \mathbb{R}^N \times \Theta$ is the solution of Eq. (6) associated with the control process $u(\cdot)$ and starting from (y, i) when $t = s$.

Let us introduce the optimal control problem under consideration in the dynamic programing form as follows. For any $(s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta$, find $\bar{u} \in U^w[s, T]$ such that

$$J(s, y, i; \bar{u}(\cdot)) = \sup_{u(\cdot) \in U^w[s,T]} J(s, y, i; u(\cdot)). \tag{8}$$

Under the following assumptions, for any $(s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta$ and $u(\cdot) \in U^w[s, T]$, the SDE in Eq. (6) admits unique solution $X(\cdot) = X_{s,y,i}(\cdot; u(\cdot))$ and the objective functional in Eq. (7) is well defined:

(A1) (U, d) is a Polish metric space, i.e., a complete separable metric space.

(A2) The maps, $f : [0, T] \times \mathbb{R}^N \times \Theta \times U \to \mathbb{R}^N$, $g : [0, T] \times \mathbb{R}^N \times \Theta \times U \to \mathbb{R}^{N\times M}$,$h : [0, T] \times \mathbb{R}^N \times \Theta \times U \times \mathbb{R}_0^K \to \mathbb{R}^{N\times K}$,$\Psi : [0, T] \times \mathbb{R}^N \times \Theta \to \mathbb{R}$, $L : [0, T] \times \mathbb{R}^N \times \Theta \times U \to \mathbb{R}$ are such that:

- for each $a \in \Theta$, $f(\cdot, \cdot, a, \cdot)$, $g(\cdot, \cdot, a, \cdot)$, $h(\cdot, \cdot, a, \cdot, \cdot)$, $\Psi(\cdot, \cdot, a)$ and $L(\cdot, \cdot, a, \cdot)$ are uniformly continuous and for each fixed $a \in \Theta$ and $k = 1, 2, ..., K$, the function defined by $\int_{\mathbb{R}_0^1} h^{(k)}(\cdot, \cdot, a, \cdot, z_k)\nu_k(dz_k)$ is also uniformly continuous;
- for each fixed $a \in \Theta$, there exists a $C > 0$ such that for $\phi(t, x, u)$=$f(t, x, a, u)$, $g(t, x, a, u)$, $\Psi(t, x, a)$ and $L(t, x, a, u)$ we have

$$|\phi(t, x, u) - \phi(t, y, u)|^2 < C\,|x - y|^2\,,$$

$$|\phi(t, 0, u)|^2 < C,$$

and for each $k = 1, 2, ..., K$, we have

$$\int_{\mathbb{R}_0^1} \left|h^{(k)}(t, x, a, u, z_k) - h^{(k)}(t, y, a, u, z_k)\right|^2 \nu_k(dz_k) < C\,|x - y|^2\,,$$

$$\int_{\mathbb{R}_0^1} \left|h^{(k)}(t, 0, a, u, z_k)\right|^2 \nu_k(dz_k) < C.$$

The value function associated with this problem is well-defined by

$$\begin{cases} V(s, y, i) = \sup\limits_{u(\cdot)\in U^w[s,T]} J(s, y, i; u(\cdot)), \\ V(T, y, i) = \Psi(T, y, i) \qquad ((s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta). \end{cases}$$

4.2 Dynamic Programing Principle and Hamilton-Jacobi-Bellman Equation

Azevedo et al. [3] obtained an extension of Bellman's optimality principle and derived the associated Hamilton-Jacobi-Bellman (HJB) equation which is a partial integro-differential equation and also its solution is the value function of the relevant optimal control problem.

Theorem 2 (Dynamic Programing Principle; [3]) *Assume that (A1)–(A2) hold. Then for any $(s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta$, we have that*

$$V(s,y,i) = \sup_{u(\cdot)\in U^w[s,T]} E\Big[\int_s^{\hat{s}} L(t, X_{s,y,i}(t;u(\cdot)), \theta_{s,i}(t), u(t))dt + V(\hat{s}, X_{s,y,i}(\hat{s};u(\cdot)), \theta_{s,i}(\hat{s}))\Big]$$

for all $\hat{s} \in [s,T]$.

In this framework, one needs the extension of Itô's formula to obtain the corresponding HJB equation. Let $tr(A)$ denote the trace of an $N \times N$ symmetric matrix A and let $\langle \cdot, \cdot \rangle$ denote the inner product in $\mathbb{R}^N$.

Lemma 1 (Itô's rule for a Markov-switching Jump-diffusion; [32]) *Suppose that* $X(t)$ *is a Markov-switching jump-diffusion process given by*

$$dX(t) = f(t, X(t_-), \theta(t_-), u(t))dt + \sum_{m=1}^{M} g(t, X(t_-), \theta(t_-), u(t))dW_m(t) + \sum_{k=1}^{K} \int_{\mathbb{R}_0^1} h^{(k)}(t, X(t_-), \theta(t_-), u(t), z_k)\tilde{J}_k(dt, dz).$$

Let $V(t,x,\theta)$ *be such that* $V(\cdot,\cdot,\theta) \in C^{1,2}([0,T]\times\mathbb{R}^N;\mathbb{R})$ *for every* $\theta \in \Theta$. *Then, we have that*

$$V(T, X(T), \theta(T)) - V(0, X(0), \theta(0)) = \int_0^T a(t, X(t), \theta(t), u(t))dt + \int_0^T b(t, X(t), \theta(t), u(t))dW(t)$$

$$+ \int_0^T c(t, X(t), \theta(t), u(t))dM(t) + \int_0^T \sum_{k=1}^{K} \int_{\mathbb{R}_0^1} d_k(t, X(t), \theta(t), u(t), z_k)\tilde{J}_k(dt, dz),$$

where

$$a(t, X, \theta, u) = V_t(t, X, \theta) + \langle V_X(t, X, \theta), f(t, X, \theta, u)\rangle + \frac{1}{2}tr(g^T(t, X, \theta, u)V_{XX}(t, X, \theta)g(t, X, \theta, u))$$

$$+ \sum_{j\in\Theta:j\neq\theta} q_{\theta j}(V(t, X, j) - V(t, X, \theta)) + \sum_{k=1}^{K} \int_{\mathbb{R}_0^1} W_k(t, X, \theta, u, V, V_X, z_k)\nu_k(dz_k)$$

with

$$W_k(t, X, \theta, u, V, V_X, z_k) = V(t, X + h^{(k)}(t, X, \theta, u, z_k), \theta) - V(t, X, \theta) - \big\langle V_X(t, X, \theta), h^{(k)}(t, X, \theta, u, z_k)\big\rangle$$

and

$$b(t, X, \theta, u) = (V_X(t, X, \theta))^T g(t, X, \theta, u),$$

$$c(t, X, \theta, u) = \sum_{j \in \Theta : j \neq \theta} (V(t, X, j) - V(t, X, \theta)),$$

$$d_k(t, X, \theta, u, z_k) = V(t, X + h^{(k)}(t, X, \theta, u, z_k), \theta) - V(t, X, \theta).$$

We now give the HJB equation associated with this optimality principle in Theorem 2.

Theorem 3 ([3]) *Suppose that the conditions (A1)–(A2) hold and that the value function V is such that $V(\cdot, \cdot, \theta) \in C^{1,2}([0, T) \times \mathbb{R}^N, \mathbb{R})$ for every $\theta \in \Theta$. Then, for each $\theta \in \Theta$, the value function $V(\cdot, \cdot, \theta)$ defined on $[0, T) \times \mathbb{R}^N$ is the solution of the Hamilton-Jacobi-Bellman equation:*

$$\begin{aligned} &V_t + \sup_{u \in U} \mathscr{H}(t, X, \theta, u, V, V_X, V_{XX}) = 0, \\ &V(T, X, \theta) = \Psi(T, X, \theta) \qquad (t, X, \theta) \in [0, T) \times \mathbb{R}^N \times \Theta, \end{aligned} \tag{9}$$

where the Hamiltonian function $\mathscr{H}(t, X, \theta, u, V, V_X, V_{XX})$ is defined by

$$\begin{aligned} \mathscr{H}(t, X, \theta, V, V_X, V_{XX}) &= L(t, X, \theta, u) + \langle V_X(t, X, \theta), f(t, X, \theta, u) \rangle \\ &+ \frac{1}{2} tr(g^T(t, X, \theta, u) V_{XX}(t, X, \theta) g(t, X, \theta, u)) + \sum_{j \in \Theta : j \neq \theta} q_{\theta j}(V(t, X, j) - V(t, X, \theta)) \\ &+ \sum_{k=1}^{K} \int_{\mathbb{R}_0^1} W_k(t, X, \theta, u, V, V_X, z_k) \nu_k(dz_k) \end{aligned}$$

and the auxiliary functions $W_k(t, X, \theta, u, V, V_X, z_k)$, $k = 1, 2, ..., K$, are defined by

$$\begin{aligned} W_k(t, X, \theta, u, V, V_X, z_k) = &V(t, X + h^{(k)}(t, X, \theta, u, z_k), \theta) \\ &- V(t, X, \theta) - \left\langle V_X(t, X, \theta), h^{(k)}(t, X, \theta, u, z_k) \right\rangle. \end{aligned}$$

Azevedo et al. [3] proved a verification theorem associated with the dynamic programing principle, Theorem 2, and the HJB equation in Eq. (9).

Theorem 4 (Verification Theorem) *Let us assume the conditions (A1)–(A2) hold and that $V(\cdot, \cdot, \theta) \in C^{1,2}([0, T) \times \mathbb{R}^N, \mathbb{R})$ for each $a \in \Theta$, $V(s, y, i)$ be a solution of the HJB equation in Eq. (9). Then, the inequality*

$$V(s, y, i) \geq J(s, y, i; u(\cdot))$$

holds for every $u(\cdot) \in U^w[s, T]$ and $(s, y, i) \in [0, T) \times \mathbb{R}^N \times \Theta$. Furthermore, an admissible pair $(\bar{X}(\cdot), \bar{\theta}(\cdot), \bar{u}(\cdot))$ is optimal for Eq. (8) if and only if the equality

$$V_t(t, \bar{X}(t), \bar{\theta}(t)) + \mathscr{H}(t, \bar{X}(t), \bar{\theta}(t), \bar{u}(t)) = 0$$

holds for a.e. $t \in [s, T]$ *and P-a.s..*

4.3 An Application to Finance

Azevedo et al. [3] studied on a consumption-investment problem for a Markov-switching jump-diffusion financial market and provided explicit optimal strategies for the power and logarithmic families of utility functions.

Within this framework, let us define a continuous-time financial market consisting of one risk-free asset, $(S_0(t) : t \in [0, T])$, and one risky asset, $(S_1(t) : t \in [0, T])$, both with the price processes evolving according to the following stochastic differential equations

$$\begin{aligned} dS_0(t) &= r(t, \theta(t_-))S_0(t_-)dt, \\ dS_1(t) &= S_1(t_-)\left\{\mu(t, \theta(t_-))dt + \sigma(t, \theta(t_-))dW(t) + \int_{\mathbb{R}_0^1} h(t, \theta(t_-), z)\tilde{J}(dt, dz)\right\}, \end{aligned}$$

with positive initial conditions $S_0(0) = s_0$ and $S_1(0) = s_1$. Let the *consumption process*, $(c(t) : t \in [0, T])$, be an $(\mathscr{F}_t)_{0\le t\le T}$-progressively measurable non-negative process satisfying the following integrability condition for the investment horizon $T > 0$:

$$\int_0^T c(t)dt < \infty \qquad a.s..$$

Let the agent's wealth allocated to the risky asset S_1, $(\alpha(t) : t \in [0, T])$ be $(\mathscr{F}_t)_{0\le t\le T}$ -progressively measurable and assume that for the fixed maximum investment horizon $T > 0$, the following integrability condition hold.

$$\int_0^T |\alpha(t)|^2 dt < \infty \qquad a.s..$$

We define the *wealth process*, $X(t)_{0\le t\le T}$ as a solution for

$$\begin{aligned} dX(t) = \{-c(t) + ((1 - \alpha(t))r(t, \theta(t_-)) + \alpha(t)\mu(t, \theta(t_-)))X(t_-)\}\, dt \qquad (10) \\ +\alpha(t)X(t_-)\left\{\sigma(t, \theta(t_-))dW(t) + \int_{\mathbb{R}_0^1} h(t, \theta(t_-), z)\tilde{J}(dt, dz)\right\}, \end{aligned}$$

with initial conditions $X(0) = x$ and $\theta(0) = a$ representing, respectively, the initial wealth and the initial state of the Markov process $\theta(\cdot)$.

Let us denote by $\mathscr{A}(x,a)$ the set of all *admissible decision strategies*, i.e., all admissible choices for the control variables $(c,\alpha) \in [0,\infty) \times [0,1]$ such that the wealth process defined by Eq. (10) is a square-integrable with respect to $dt \times dP$ over $[0,T] \times \Omega$. The consumption-investment problem is to find the consumption and investment strategies, $(c,\alpha) \in \mathscr{A}(x,a)$, which maximize the expected utility

$$J(s,y,i;c(\cdot),\alpha(\cdot)) = E\left[\int_s^T U(t,c(t),\theta_{s,i}(t))dt + \Psi(T, X_{s,y,i}(T;c(\cdot),\alpha(\cdot)),\theta_{s,i}(T))\right], \tag{11}$$

where $(X_{s,y,i}(t;c(\cdot),\alpha(\cdot)),\theta_{s,i}(t)) \in \mathbb{R} \times \Theta$ is the solution of Eq. (10) associated with the strategies $c(\cdot),\alpha(\cdot)$ and starting from (y,i) when $t = s$.

Let us present the optimal consumption-investment strategies for utility functions belonging to the following class of power utilities:

$$U(t,c,a) = e^{-\rho t}\frac{c^{\gamma_a}}{\gamma_a}, \qquad \Psi(t,x,a) = e^{-\rho t}\frac{x^{\gamma_a}}{\gamma_a}, \tag{12}$$

where $\gamma_a \in (0,1)$ is the risk aversion coefficient associated with the state of the Markov process $\theta(t) = a \in \Theta$, and $\rho > 0$ is the discount rate. These families of utility functions have a constant coefficient of relative risk aversion or Arrow-Pratt-De Finetti measure of relative risk aversion which make the computation of closed form solutions for HJB equations easier in this special case.

Before providing the precise theorem, let us introduce the function $F : [0,1] \times [0,T] \times \Theta \to \mathbb{R}$ given by

$$\begin{aligned} F(\alpha;t,a) &= \gamma_a[r(t,a) + \alpha(\mu(t,a) - r(t,a)) - \frac{1}{2}(1-\gamma_a)\alpha^2\sigma^2(t,a)] \\ &+ \int_{\mathbb{R}_0^1}[(1+\alpha h(t,a,z))^{\gamma_a} - 1 - \gamma_a\alpha h(t,a,z)]\nu(dz). \end{aligned}$$

Theorem 5 ([3]). *The maximum expected utility associated with Eq. (11) and the discounted utility functions of Eq. (12) is given by following value function*

$$V(t,x,a) = \xi_a(t)\frac{x^{\gamma_a}}{\gamma_a},$$

the corresponding optimal strategies are of the form

$$c^*(t,x,a) = x(e^{\rho t}\xi_a(t))^{-1/(1-\gamma_a)}$$

and

$$\alpha^*(t,a) = \begin{cases} 1, & \text{if } \mu(t,a) > r(t,a) \text{ and } F'(1;t,a) \geq 0, \\ \hat{\alpha}(t,a), & \text{if } \mu(t,a) > r(t,a) \text{ and } F'(1;t,a) < 0, \\ 0, & \text{if } \mu(t,a) \leq r(t,a), \end{cases}$$

where $\hat{\alpha}(t,a)$ *is the unique solution of* $F'(\alpha;t,a) = 0$ *in* $(0,1)$ *and* $\xi_a(t)$*,* $a \in \Theta$*, are solutions of the following coupled ordinary differential equations terminal value problem*

$$\xi_a'(t) + (1-\gamma_a)e^{-\rho t/(1-\gamma_a)}\xi_a(t)^{-\gamma_a/(1-\gamma_a)} + F(\alpha^*(t,a);t,a)\xi_a(t) + \sum_{j\in\Theta:j\neq a} q_{aj}(\xi_j(t)-\xi_a(t)) = 0, \qquad \xi_a(T) = e^{-\rho T}.$$

5 Conclusion and Outlook

SHSJs are natural, powerful and efficient candidates to model abrupt changes in financial markets as a consequence of their heterogenous nature. Whereas some changes in the financial market may be transitory, often the new behavior of the market persists for many periods, e.g., recessions versus expansions. In some instances, a discrete shift from one regime to another may result from a change in economic policy, e.g., a shift in a monetary or an exchange rate regime, and in some others, it may be activated by a major event, such as the bankruptcy of Lehman Brothers in September 2008, or the 1973 oil crisis. Regime switching models can capture not only the sudden changes of behavior of financial markets but also the new dynamics and fundamentals persist for several periods after a change. Since an investor may have to compensate large costs of ignoring the regimes, an intuitive and expected question is centered on the existence of an optimal portfolio to hedge against the risk of regime changes, and the other one is how to determine the portfolios which should be optimally held in each regime. At this point, the crucial role of stochastic optimal control theory in finance and economics becomes highlighted.

Within this framework, dynamic programing techniques for Markov-switching jump diffusions have been considered by various researchers [3, 13–15, 35, 36]. On the other hand, in Markovian models, an investor has to make his investment decisions based only on the current information and will have no chance to take into account the history of the stock prices or wealth process. However, in real-world problems, the investor tends to look at the historical performance of the risky asset or his portfolio which may depend on their own past, before he makes his investment decision. Therefore, in such cases, where we need memory in the model, the theory of SDEs with time delay, *SDDE*, becomes an excellent and applicable tool in stochastic optimal control theory.

From this perspective, a comprehensive theory on SDEs with memory can be found in the monograph by Mohammed [27, 28]. Also stochastic optimal control theory with time delay has been presented in their detailed monograph by Kolmanovskii and Shaikhet [24]. Several other works have been addressed for optimal control of SDDE in the framework of the maximum principle by Elsanosi et al. [16], Øksendal and Sulem [29] and Øksendal et al. [31], in addition from the HJB equations' point of view by David [11] and Federico [17]. Although SHSs and SHSJs have been considered by many authors, there have been few works on SHSs and SHSJs with delay. To the best of our knowledge, Mao et al. [26] studied exponential stability of SDDE with Markovian switching.

From a different point of view, Temocin and Weber [33] have developed numerical approximation methods for solving the optimal control problems associated with the controlled autonomous stochastic hybrid systems with jumps. The approximation has been done on the basis of Markov decision processes that preserve local consistency with the original solution. Further development of numerical implementations of SHSJs is another area of future work for research with potential applications to finance and economics.

In addition to regime switches, our study is also applicable for sudden paradigm changes. A paradigm is our perception of reality, our view of the world and from time to time it may make relatively sudden shifts to radically new paradigms which means to have a sudden change in perception, a sudden change in point of view, of how we see things. Hence, the crisis generated as a consequence of such gaps creates cultural as well as societal transformations and they may undergo "switches", as far as they can be represented by stochastic dynamics which are investigated in this chapter. Let us mention that, in the real world, cultural, societal, economical and financial developments are closely related with each other. For example, globalization opens up economic opportunities while making states and their people increasingly vulnerable to destabilizing impacts from beyond national borders. In this respect, the present chapter opens a new and wider view on the world of today and tomorrow.

References

1. Abate, A., Prandini, M., Lygeros, J., Sastry, S.: Approximation of general stochastic Hybrid systems by switching diffusions with random hybrid jumps. hybrid systems: computation and control. Lect. Notes Comput. Sci. **4981**, 598–601 (2008)
2. Amonlirdviman, K., Khare, N.A., Tree, D.R.P., Chen, W.S., Axelrod, J.D., Tomlin, C.J.: Mathematical modeling of planar cell polarity to understand domineering nonautonomy. Sci. **307**(5708), 423–426 (2005)
3. Azevedo, N., Pinheiro, D., Weber, G.W.: Dynamic programming for a Markov-switching jumpdiffusion. J. Comput. Appl. Math. **267**, 1–19 (2014)
4. Balluchi, A., Benvenuti, L., Benedetto, M.D.D., Miconi, G.M., Pozzi, U., Villa, T., Wong-Toi, H., Sangiovanni-Vincentelli, A.L.: Maximal safe set computation for idle speed control of an automotive engine. In: Lynch, N., Krogh, B.H. (eds.) Hybrid Systems: Computation and Control. Lecture Notes in Computer Science, pp. 32–44. Springer, Berlin (2000)
5. Bellman R.E.: On the theory of dynamic programming. Proc. Natl. Acad. Sci. **38**, 716–719. USA (1952)
6. Bellman, R.E.: Dynamic programming and stochastic control process. Inf. Control **1**, 228–239 (1958)

7. Bensoussan, A., Menaldi, J.L.: Stochastic hybrid control. J. Math. Anal. Appl. **249**, 261–288 (2000)
8. Blom H.A.P.: Stochastic hybrid processes with hybrid jumps. In: Engell, S., Gueguen, H., Zaytoon, J. (eds.) Proceedings of the IFAC Conference Analysis and Design of Hybrid Systems (ADHS 2003), Saint-Malo, Brittany, France, 16–18 June 2003
9. Blom H.A.P., Bakker G.J., Everdij M.H.C., van der Park M.N.J.: Stochastic analysis background of accident risk assessment for Air Traffic Management. Hybridge Report D2.2. National Aerospace Laboratory NLR. Available via DIALOG. http://www.nlr.nl/public/hosted-sites/hybridge (2003)
10. Bujorianu M.L., Lygeros J., Glover W., Pola G.: A stochastic hybrid system modeling framework. Technical Report WP1, Deliverable D1.2, Hybridge (2002)
11. David D.: Optimal control of stochastic delayed systems with jumps. Preprint (2008)
12. Davis, M.H.A., Vellekoop, M.H.: Permanent health insurance: a case study in piecewise-deterministic markov modelling. Mitteilungen der Schweiz. Vereinigung der Versicherungsmathematiker **2**, 177–212 (1995)
13. Elliott, R.J., Chan, L., Siu, T.: Option pricing and Esscher transform under regime switching. Ann. Financ. **1**, 423–432 (2005)
14. Elliott, R.J., Siu, T.: On risk minimizing portfolios under a markovian regime switching Black-Scholes economy. Ann. Oper. Res. **176**, 271–291 (2010)
15. Elliott, R.J., Zhang, X., Siu, T.K.: A stochastic maximum principle for a Markov regime-switching jump-diffusion model and its application to finance. SIAM J. Control Optim. **50**(2), 964–990 (2012)
16. Elsanosi, I., Øksendal, B., Sulem, A.: Some solvable stochastic control problems with delay. Stoch. **71**(1–2), 69–89 (2000)
17. Federico, S.: A stochastic control problem with delay arising in a pension fund model. Financ. Stoch. **15**(3), 421–459 (2011)
18. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions, 2nd edn. Springer, New York (2006)
19. Framstad, N.C., Øksendal, B., Sulem, A.: Sufficient stochastic maximum principle for the optimal control of jump diffusions and applications to finance. J. Optim. Theory Appl. **121**, 77–98 (2004)
20. Ghosh, M.K., Arapostathis, A., Marcus, S.I.: Optimal control of switching diffusions with application to flexible manufacturing systems. SIAM J. Control Optim. **31**(5), 1183–1204 (1993)
21. Ghosh, M.K., Arapostathis, A., Marcus, S.I.: Ergodic control of switching diffusions. SIAM J. Control Optim. **35**(6), 1952–1988 (1997)
22. Ghosh M.K., Bagchi A.: Modeling stochastic hybrid systems. In: 21st IFIP TC7 Conference on System Modelling and Optimization (2003)
23. Hu, J., Lygeros, J., Sastry, S.: Towards a theory of stochastic hybrid systems. In: Lynch, N., Krogh, B.H. (eds.) Hybrid Systems: Computation and Control, number 1790. LNCS, pp. 160–173. Springer, Berlin (2000)
24. Kolmanovskii V.B., Shaikhet L.E.: Control of Systems with Aftereffect. Translations of Mathematical Monographs, vol. 157. American Mathematical Society, Providence (1996)
25. Krystul J., Blom H.A.P.: Generalized stochastic hybrid processes as strong solutions of stochastic differential equations. Hybridge report D2.3 Available via DIALOG. http://www.nlr.nl/public/hosted-sites/hybridge/ (2005)
26. Mao, X., Matasov, A., Piunovskiy, A.B.: Stochastic differential delay equations with Markovian switching. Bernoulli **6**(1), 73–90 (2000)
27. Mohammed S.: Stochastic Functional Differential Equations. Pitman (1984)
28. Mohammed S.: Stochastic differential systems with memory. Theory, examples and applications. In: Decreusefond L. et al. (eds.) Stochastic Analysis and Related Topics VI. The Geilo Workshop 1996, pp. 1–77. Birkauser, Switzerland (1998)
29. Øksendal, B., Sulem, A.: A maximum principle for optimal control of stochastic systems with delay with applications to finance. In: Menaldi, J.L., Rofman, E., Sulem, A. (eds.) Optimal

Control and PDE. Essays in Honour of Alain Bensoussan, pp. 64–79. IOS Press, Amsterdam (2001)

30. Øksendal B., Sulem A.: Applied Stochastic Control of Jump Diffusions. Springer (2005)
31. Øksendal, B., Sulem, A., Zhang, T.: Optimal control of stochastic delay equations and time-advanced backward stochastic differential equations. Adv. Appl. Probab. **43**(2), 572–596 (2011)
32. Protter, P.E.: Stochastic Integration and Differential Equations, 2nd edn. Springer, New York (2005)
33. Temocin, B.Z., Weber, G.W.: Optimal control of stochastic hybrid system with jumps: a numerical approximation. J. Comput. Appl. Math. **259**, 443–451 (2014)
34. Watkins O.J., Lygeros J.: Safety relevant operational cases in ATM. Technical Report WP, Deliverable D1.1, Hybridge (2002)
35. Zhang, Q.: Stock trading: an optimal selling rule. SIAM J. Control Optim. **40**, 64–87 (2001)
36. Zhang, Q., Yin, G.: Nearly-optimal asset allocation in hybrid stock investment models. J. Optim. Theory Appl. **121**, 419–444 (2004)

State-Dependent Impulsive Neural Networks

Mustafa Şaylı and Enes Yılmaz

Abstract In this chapter, we address a new model of neural networks related to the discontinuity phenomena which is called state-dependent impulsive neural networks. By means of B-equivalence method, we reduce these networks to a fix time impulsive neural networks system. In the first part of this study, sufficient conditions for existence and uniqueness of exponentially stable almost periodic solution for recurrent neural networks are investigated. In the second part, sufficient conditions ensuring the existence, uniqueness and global robust asymptotic stability of the equilibrium point for a more general class of bidirectional associative memory (BAM) neural networks are obtained by employing an appropriate Lyapunov function and linear matrix inequality (LMI). Finally, an illustrative example is given to show the effectiveness of the theoretical results.

Keywords Neural networks · State-dependent impulsive systems · Stability · Linear matrix inequality

1 Introduction

Neural networks have many important applications in pattern recognition, signal processing, associative memory, and optimization problems. During recent years, with the help of hardware implementation, neural dynamical methods for solving optimization problems have received considerable interest. The difficulty of dynamic optimization is significantly amplified when the optimal solutions have to be obtained in real time, especially in the presence of uncertainty. In such applica-

M. Şaylı
School of Mathematical Sciences, University of Nottingham,
Nottingham NG7 2RD, UK
e-mail: mustafasaylitr@gmail.com

E. Yılmaz (✉)
College of Computing and Digital Media, School of Computing, DePaul University,
243 S Wabash Ave, Chicago, IL 60604, USA
e-mail: eyilmaz@depaul.edu

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_19

tions, compared with traditional numerical optimization algorithms, neurodynamic optimization approaches based on recurrent neural networks have several unique advantages. Recurrent neural networks can be physically implemented in designated hardware/firmware, such as digit signal processor(DSP), optical chips, graphic processing units (GPU), very-large-scale integration (VLSI) reconfigurable analog chips, field programmable gate array (FPGA) [45, 52, 76]. In addition, neural networks, especially Hopfield type neural networks, can be considered as a nonlinear associative memory or content-addressable memory (CAM). In the application of these networks, an important property of the CAM is the ability to retrieve a stored pattern, given a reasonable subset of the information content of that pattern [28]. Apparently, since these networks have ability to learn, one can easily consider learning theory related to an unsupervised Hebbian-type learning mechanism [20, 24, 26, 35]. All of these applications tediously depend on dynamical behaviors of the network and require that the equilibrium point of the model is globally asymptotically stable. There are many studies on the global stability analysis and other dynamical behaviors, like periodicity/almost periodicity, of neural networks [13, 16, 31, 57, 58, 65, 66, 74, 78, 79, 84] and references therein.

Impulsive neural networks (see, for example [1, 6–8, 15, 19, 21–23, 27, 30, 37, 40, 41, 43, 44, 48–51, 59, 63, 67, 70, 72, 73, 80, 81, 83]) have been enormously developed issuing from the fact that in implementation of electronic networks, the state of the networks is subject to instantaneous perturbations and experiences abrupt change at certain moments, which may be caused by switching phenomenon, frequency change or other sudden noises. Furthermore, the dynamics of quasi-active dendrites with active spines is described by a system of point hot-spots (with an integrate-and-fire process), see [18, 64] for more details. On the other hand, studies on neural dynamical systems not only involve stability and periodicity, but also involve other dynamic behaviors such as almost periodicity, chaos and bifurcation. If one considers long-term dynamical behaviors, the periodic parameters often turn out to experience certain perturbations, that is, parameters become periodic up to a small error. Thus, almost periodic oscillatory behavior is considered to be more accordant with reality. Although it is of great importance in real life applications, the generalization to almost periodicity has been rarely studied in the literature; see [9, 10, 17, 25, 38, 51, 59, 67–69, 83]. Moreover, in practical implementation of neural networks, the stability of networks can often be destroyed by its compulsory uncertainty issuing from the existence of modeling errors, external disturbance and parameter fluctuations. In addition, several studies with interesting results examining robust stability analysis of neural networks were published in [12, 14, 39, 42, 46, 55, 56, 82]. Hence, robustness of the designed network should be considered. In the light of above discussion, it is necessary to consider both impulsive phenomena and robustness of the neural networks (see, for example [42, 77, 82]).

The aim of defining this new class is that the moments of discontinuity θ_k are arbitrary in $\mathbb{R}$. In the real world problems, the impulses of many systems do not occur at fixed times but depends on the states of the systems, for example, some circuit control systems, saving rate control systems and population control systems and so on. These types of systems are called state-dependent impulsive differential systems or impulsive systems with variable-time impulses. In the current study, different from the most existing studies, we discuss almost periodicity as well as robustness of the neural networks having impulse times at the hyper surfaces $\Gamma_k : t = \theta_k + \tau_k(x), k \in \mathbb{Z}$, not on the planes $t = \theta_k$. For a more detailed discussion about real world applications for state-dependent impulsive systems, we refer the reader to books written by M. Akhmet and T. Yang (See [3, 71]). Therefore, consideration on the system with non-fix moments of impulses is necessary and more general than the fixed time impulsive systems, which results in more theoretical and technical challenges. This is due to the fact that, we know the difference of any two solutions is again a solution, but all of these solutions have different discontinuity time and dynamical behaviors should not be same. So, simple transformations are not allowed for state-dependent impulsive systems. To solve the problem, we should develop the technique of the reduction of the considered system to system with fixed moments of impulses. That is, B-equivalence method, which was studied for bounded domain in the phase space [2–5].

The rest of the chapter is organized as follows. In Sect. 2, sufficient conditions for existence and uniqueness of exponentially stable almost periodic solution for impulsive recurrent neural networks with variable moments of time are investigated by virtue of contraction mapping principle and Gronwall-Bellman lemma. In Sect. 3, global robust asymptotic stability of the equilibrium point for a more general class of bidirectional associative memory (BAM) neural networks with variable time of impulses is studied. Sufficient conditions ensuring the existence, uniqueness and global robust asymptotic stability of the equilibrium point are obtained by employing an appropriate Lyapunov function and linear matrix inequality (LMI). Finally, a conclusion and discussion is given in Sect. 4.

2 Almost Periodicitiy of State-Dependent Impulsive Neural Networks

In this part of following chapter, we will investigate sufficient conditions for existence and uniqueness of exponentially stable almost periodic solution for impulsive recurrent neural networks with variable moments of time. As a model, we analyze recurrent neural networks, however obtained results can be effectively applied to almost all problems in neural networks models.

2.1 Model Description and Preliminaries

Let $\mathbb{Z}$ and $\mathbb{R}$ be the sets of integers and real numbers. Consider the following impulsive recurrent neural networks with variable moments of time:

$$\begin{aligned} x_i'(t) &= -a_i(t)x_i(t) + \sum_{j=1}^{m} b_{ij}(t) f_j(x_j(t)) + c_i(t), \\ \Delta x_i \mid_{t=\theta_k+\tau_k(x)} &= d_{ik}x_i + I_{ik}(x), \end{aligned} \tag{1}$$

where $a_i(t) > 0, \;\; i = 1, 2, \ldots, m, \;\; k \in \mathbb{Z}, \;\; x \in \mathbb{R}^m, t \in \mathbb{R}, \; \{d_{ik}\}$ is a bounded sequence such that $(1 + d_{ik}) \neq 0, \; i = 1, 2, \cdots, m, \; k \in \mathbb{Z}, \tau_k(x)$ are positive real valued continuous functions defined on $\mathbb{R}^m, \; k \in \mathbb{Z}$. Moreover, the sequence θ_k satisfies the following condition $\theta_k < \theta_{k+1}, \; |\theta_k| \to +\infty \;\;$ as $\;\; |\mathrm{k}| \to \infty$.

In system (1), $x_i(t)$ denotes the membrane potential of the unit i at time t; the continuous functions $f_j(.)$ represent the measures of activation to its incoming potentials of the unit j at time t; b_{ij} corresponds to the synaptic connection weight of the unit j on the unit i; c_i signifies the external bias or input from outside the network to the unit i; a_i is the rate with which the ith unit will reset its potential to the resting state in isolation when it is disconnected from the network and external inputs. It will be assumed that $a_i, b_{ij}, c_i, I_{ik} : \mathbb{R}^m \to \mathbb{R}^m$ are continuous functions.

The following assumptions will be needed throughout the section:

(A1) there exists a Lipschitz constant $\ell > 0$ such that

$$|\tau_k(x) - \tau_k(y)| + |f_i(x) - f_i(y)| + |I_{ik}(x) - I_{ik}(y)| \leq \ell |x - y|$$

and $|\tau_k(x)| < \ell$ for all $x, y \in \mathbb{R}^m, \; i = 1, 2, \ldots, m, \; k \in \mathbb{Z}$;

(A2) there exists a positive number $\theta \in \mathbb{R}$ such that $\theta_{k+1} - \theta_k \geq \theta$ holds for all $k \in \mathbb{Z}$ and the surfaces of discontinuity $\Gamma_k : t = \theta_k + \tau_k(x), \; k \in \mathbb{Z}$ satisfy the following conditions:

$$\theta_k + \tau_k(x) < \theta_{k+1} + \tau_{k+1}(x), \; |\theta_k| \to +\infty \quad \text{as} \quad |\mathrm{k}| \to \infty,$$

$$\tau_k((E + D_k)x + I_k(x)) \leq \tau_k(x), \; x \in \mathbb{R}^m,$$

where E is an $m \times m$ identity matrix and

$$D_k = diag(d_{1k}, \cdots, d_{mk}) = \begin{pmatrix} d_{1k} & 0 & \ldots & 0 \\ 0 & d_{2k} & \ldots & 0 \\ \ldots & & & \\ 0 & 0 & \ldots & d_{mk} \end{pmatrix} \text{and } \mathrm{I_k} = \begin{pmatrix} I_{1k} \\ I_{2k} \\ \ldots \\ I_{mk} \end{pmatrix}.$$

For the sake of convenience, we adopt the following notations in the sequel:

$$k_1 = \max_{1 \le i \le m} \sup_{t \in \mathbb{R}} \left(|a_i(t)| + \ell \sum_{j=1}^{m} |b_{ji}(t)| \right) < +\infty,$$

$$k_2 = \max_{1 \le i \le m} \sup_{t \in \mathbb{R}} \left(\sum_{j=1}^{m} |b_{ji}(t)||f_i(0)| + |c_i(t)| \right) < +\infty,$$

$$k_3 = \max_{1 \le i \le m} \sup_{t \in \mathbb{R}} \left(\sum_{j=1}^{m} |b_{ji}(t)| \right) < +\infty, \ k_4 = \max_{k \ge 1} \Big(|J_{ik}(0)| \Big) < +\infty.$$

Now, we need the following assumption to avoid beating phenomena.

(A3) $\ell(k_1 h + k_2) < 1$.

In order to find a more detailed discussion on the condition (A3), we refer the interested readers to the Lemma 5.3.1 in the book [3].

Definition 1 A function $x(t) : \mathbb{R}^m \to \mathbb{R}^m$ is said to be a solution of the system (1), if

(i) $x(t)$ satisfies system (1) for all $t \in \mathbb{R}$
(ii) $x(t)$ is continuous everywhere expect on the surfaces $\Gamma_k : t_k = \theta_k + \tau_k(x)$ and left continuous at the planes $t = t_k$, and the right limit $x(t_k^+)$ exists for $k \in \mathbb{Z}$.

From local existence theorem (Theorem 5.2.1 in [3]), a solution of (1) exists. By virtue of Theorem 5.3.1 in [3] and assumptions (A2)–(A3), every solution $x(t),\ ||x(t)|| \le h$ of (1) intersects each surface of discontinuity $\Gamma_k : t = \theta_k + \tau_k(x),\ k \in \mathbb{Z}$, at most once. Furthermore, by the proof of Theorem 5.2.4 in [3], continuation of solutions of the ordinary differential equation

$$x_i'(t) = -a_i(t)x_i(t) + \sum_{j=1}^{m} b_{ij}(t) f_j(x_j(t)) + c_i(t) \tag{2}$$

and the condition $|\theta_k| \to +\infty$ as $|\mathrm{k}| \to \infty$, one can find that every solution $x(t) = x(t, t_0, x^0),\ (t_0, x^0) \in \mathbb{R} \times \mathbb{R}^m$, of (1) is continuable on $\mathbb{R}$. That is to say, the interval of existence is whole real line.

2.2 *Reduced B-Equivalence Systems*

A difficulty at the investigation of a system (1) is that the discontinuities of distinct solutions are not, in general the same. To investigate the asymptotic properties of solutions of Eq. (1), we introduce the following concepts. In what follows, we give the techniques of B-topology and B-equivalence which were introduced and developed in [3] for the systems of differential equations with variable moments of time. For detailed discussion, we refer to reader to the book [3].

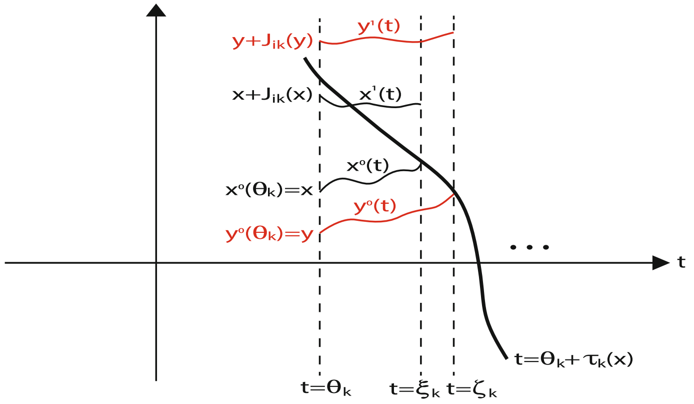


Fig. 1 The procedure of the construction of the map J_{ik}

For a fix $k \in \mathbb{Z}$, let $x^0(t) = x(t, \theta_k, x^0)$ be a solution of the system of ordinary differential equations (2). Denote by $t = \xi_k$ the time when the solution of (2) intersects the surface of discontinuity $\Gamma_k : t = \theta_k + \tau_k(x(\xi_k)),\ k \in \mathbb{Z}$. Suppose that $x^1(t) = x(t, \xi_k, (E + D_k)x^0(\theta_k) + I_{ik}(x^0(\xi_k)))$ is also a solution of (2). Next, we define a mapping $J_{ik}(x) : \mathbb{R}^m \to \mathbb{R}^m$ such that $J_{ik}(x) = x^1(\theta_k) - (E + D_k)x$ (Fig. 1 illustrates the procedure of the contraction of the map J_{ik}.) and construct a system of impulsive differential equations with fixed moments, which has the form

$$\begin{aligned} y_i'(t) &= -a_i(t)y_i(t) + \sum_{j=1}^{m} b_{ij}(t)f_j(y_j(t)) + c_i(t), \\ \Delta y_i\ |_{t=\theta_k} &= d_{ik}y_i + J_{ik}(y), \end{aligned} \tag{3}$$

where $a_i(t) > 0,\ i = 1, 2, \ldots, m,\ k \in \mathbb{Z}$.

We denote $\mathscr{PC}(J; \mathbb{R}^m),\ J \subset \mathbb{R}$, the space of all piecewise continuous functions $\varphi : J \to \mathbb{R}^m$ with points of discontinuity of the first kind $\theta_k,\ k \in \mathbb{Z}$ and which are continuous from the left.

Let $x(t)$ be a solution of Eq. (1) on $\mathscr{U}$ ($\mathscr{U}$ can be an interval, a real half-line, or the real line $\mathbb{R}$).

Definition 2 A solution $y(t)$ of (3) is said to be in the ε-neighborhood of a solution $x(t)$ if:

(i) the measure of the symmetrical difference between the domains of existence of these solutions does not exceed ε;

(ii) discontinuity points of $y(t)$ are in ε-neighborhoods of discontinuity points of $x(t)$;

(iii) for all $t \in \mathscr{U}$ outside of ε-neighborhoods of discontinuity points of $x(t)$ the inequality $\|x(t) - y(t)\| < \varepsilon$ holds.

The topology defined by ε- neighborhoods of piecewise continuous solutions will be called the B-topology. It is easily seen that it is Hausdorff topology. Topologies and metrics for spaces of discontinuous functions were introduced and developed in [2, 4, 32].

For any $u, v \in \mathbb{R}$ we define the oriented interval $\widehat{[u, v]}$ as

$$\widehat{[u, v]} = \left\{ \begin{array}{l} [u, v]\,, \;\; \text{if } u \le v \\ [v, u]\,, \text{ otherwise} \end{array} \right\}. \tag{4}$$

Definition 3 Systems (1) and (3) are said to be B-equivalent, if for any solution $x(t)$ of (1) defined on an interval $\mathscr{U}$ with the discontinuity points $\xi_k,\ k \in \mathbb{Z}$, there exists a solution $y(t)$ of system (3) satisfying

$$x(t) = y(t),\ t \in \mathbb{R}/\bigcup_{k\in\mathbb{Z}} \widehat{(\xi_k, \theta_k]}. \tag{5}$$

In particular,

$$x(\theta_k) = y(\theta_k+),\ x(\xi_k) = y(\xi_k) \text{ if } \theta_k > \xi_k \tag{6}$$

$$x(\theta_k) = y(\theta_k),\ x(\xi_k+) = y(\xi_k) \text{ if } \theta_k < \xi_k \tag{7}$$

where $y(\theta_k+)$ and $x(\xi_k+)$ are the right limits of $y(t)$ and $x(t)$ at the points θ_k and ξ_k, respectively. Conversely, for each solution $y(t)$ of Eq. (3), there exists a solution $x(t)$ of system (1), which satisfies the conditions (5)–(7).

The proof of following lemma is quite similar to that of Theorem 5.8.1 in [3], so we omit it here.

Lemma 1 *There are mappings $J_k(y) : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, $k \in \mathbb{Z}$, such that corresponding to each solution $x(t)$ of Eq. (1), there is a solution $y(t)$ of the system (3) satisfying $x(t) = y(t)$ if $t \in \mathbb{R}\backslash \bigcup_{k\in\mathbb{Z}} \widehat{(\xi_k, \theta_k]}$. Moreover, the functions $J_{ik}(y)$ satisfy the inequality*

$$||J_k(x) - J_k(y)|| \le \ell k(\ell)||x - y||, \tag{8}$$

where $k(\ell) = k(\ell, h)$ is a bounded function, uniformly with respect to $k \in \mathbb{Z}$ for all $x, y \in \mathbb{R}^m$, such that $||x|| \le h$ and $||y|| \le h$.

2.3 Existence and Uniqueness of Exponentially Stable Almost Periodic Solution

In this section, we analyze the almost periodic (a. p.) systems. Assume that the system (1) is a. p., i.e., the functions $a_i(t), b_{ij}(t), c_i(t)$ and $f_j(x_j(t))$ are Bohr a. p. as functions of t, the matrix sequences D_k, and sequences $I_k(x), \tau_k(x)$ are a. p.

with respect to k, uniformly with respect to x, $||x|| \leq h$. The sequences $\{\theta_k^j\}$ where $\theta_k^j = \theta_{k+j} - \theta_k$ are uniformly a.p.(equipotentially a. p.) with respect to k.

Denote the set of all sequences by

$$\mathscr{C} = \left\{ \{\theta_k\} : \theta_k \in \mathbb{R},\ \theta_k < \theta_{k+1},\ k \in \mathbb{Z},\ \lim_{k \to \pm\infty} \theta_k = \pm\infty \right\}$$

and adopt the following essential definitions and lemmas for almost periodicity.

Definition 4 ([53]) The set of sequences $\{\theta_k^j\}$, $\theta_k^j = \theta_{k+j} - \theta_k$, $k \in \mathbb{Z}$, $j \in \mathbb{Z}$, $\{\theta_k\} \in \mathscr{C}$ is called uniformly almost periodic if for arbitrary $\varepsilon > 0$ there exists relatively dense set of ε-almost periods common for any sequences.

Definition 5 ([53]) A piecewise continuous function $\varphi : \mathbb{R} \to \mathbb{R}^m$ with discontinuity of first kind at the points θ_k is called almost periodic, if

(a) the set of sequences $\{\theta_k^j\}$, $\theta_k^j = \theta_{k+j} - \theta_k$, $k \in \mathbb{Z}$, $j \in \mathbb{Z}$, $\{\theta_k\} \in \mathscr{C}$ is uniformly almost periodic.

(b) for any $\varepsilon > 0$ there exists a real number $\delta > 0$ such that if the points t_1 and t_2 belong to the same interval of continuity of $\varphi(t)$ and satisfy the inequality $|t_1 - t_2| < \delta$, then $|\varphi(t_1) - \varphi(t_2)| < \varepsilon$.

(c) for any $\varepsilon > 0$ there exists a relatively dense set T such that if $r \in T$, then $|\varphi(t + r) - \varphi(t)| < \varepsilon$ for all $t \in \mathbb{R}$ satisfying the condition $|t - \theta_k| > \varepsilon$, $k \in \mathbb{Z}$.

Lemma 2 ([53]) *Assume that an a. p. system (1) satisfies all foregoing conditions. Then, for each $\varepsilon > 0$ there exist ε_1, $0 < \varepsilon_1 < \varepsilon$ and relatively dense sets $r \in T$ of real numbers and $q \in \mathbb{Q}$ of rational numbers, such that the following inequalities are valid:*

- $|a_i(t + r) - a(t)| < \varepsilon,\ t \in \mathbb{R},\ |t - \theta_k| > \varepsilon,\ i = 1, \ldots, m;$
- $|b_{ij}(t + r) - b_{ij}(t)| < \varepsilon,\ t \in \mathbb{R},\ |t - \theta_k| > \varepsilon,\ i, j = 1, \ldots, m;$
- $|c_i(t + r) - c_i(t)| < \varepsilon,\ t \in \mathbb{R},\ |t - \theta_k| > \varepsilon,\ i = 1, \ldots, m;$
- $\sup_{|x| \leq h} |f_j(x_j(t + r)) - f_j(x_j(t))| < \varepsilon,\ t \in \mathbb{R},\ |t - \theta_k| > \varepsilon,\ j = 1, \ldots, m;$
- $|D_{k+q} - D_k| < \varepsilon,\ k \in \mathbb{Z};$
- $\sup_{|x| \leq h} |I_{k+q}(x) - I_k(x)| < \varepsilon,\ k \in \mathbb{Z};$
- $|\overline{\theta}_k^q - r| < \varepsilon_1,\ k \in \mathbb{Z};$
- $\sup_{|x| \leq h} |\tau_{k+q}(x) - \tau_k(x)| < \varepsilon,\ t \in \mathbb{R},\ k \in \mathbb{Z}.$

The proof is similar to that of Lemma 5 in paper [4] and thus we omit it here.

Lemma 3 *If the Eq. (1) is a. p., then the mapping $J_{ik}(x)$ is a. p. with respect to k, uniformly with respect to x, $|x| \leq h$.*

Let $X(t) = diag(X_1(t), \cdots, X_m(t))$ be a fundamental matrix solution of the associated system (1),

$$\begin{aligned} x_i'(t) &= -a_i(t)x_i(t) \\ \Delta x_i \mid_{t=\theta_k} &= d_{ik}x_i(t), \end{aligned} \tag{9}$$

where $a_i(t) > 0,\ i = 1, 2, \ldots, m,\ k \in \mathbb{Z}$ such that $X(0) = E$. Denote by $X_i(t, s) = X_i(t)X_i^{-1}(s),\ t, s \in \mathbb{R},\ i = 1, 2, \cdots, m$ the transition matrix of (9). For $X_i(t, s)$, assume that the following inequality holds:

$$|X_i(t, s)| \leq Ke^{-\lambda(t-s)},\ t \geq s$$

where K and λ are positive real numbers, under the condition

$$\lambda = \inf_{t\in\mathbb{R},\ 1\leq i\leq m} |a_i(t)| > 0.$$

Now, let us give the following assertions, which will be important in the main theorem. We omit the proofs since the proofs are similar to that of Lemma 36 and Lemma 37 in [54].

Lemma 4 *If the transition matrix $X_i(t, s)$ satisfies the inequality*

$$|X_i(t, s)| \leq Ke^{-\lambda(t-s)},\ t \geq s,$$

where K and λ are positive real numbers, then for any $\varepsilon > 0,\ t \in \mathbb{R},\ s \in \mathbb{R},\ t \geq s,\ |t - \theta_k| > \varepsilon,\ |s - \theta_k| > \varepsilon,\ k \in \mathbb{Z}$, there exists a relatively dense set T of almost periods, Γ, such that, for $r \in \Gamma$, we have

$$|X_i(t + r, s + r) - X_i(t, s)| \leq \varepsilon\Gamma e^{\frac{-\lambda}{2}(t-s)},$$

where Γ is a positive constant.

Lemma 5 *If $\varphi(t)$ is an a. p. function and $\inf\limits_{k\in\mathbb{Z}} \theta_k' = \theta > 0$, then $\{\varphi(t_k)\}$ is an a. p. sequence.*

Now, by virtue of contraction mapping principle and Gronwall-Bellman lemma, we obtain sufficient conditions of a unique exponentially stable almost periodic solution for system (1).

Let $\varepsilon > 0$ be such that $\lambda = \lambda(\varepsilon) > 0$ and

$$\mathscr{M} = \frac{k_2(e^{\lambda\theta-1}) + \lambda k_4 e^{2\lambda\theta}}{(e^{\lambda\theta-1})(\lambda - \ell K k_3) - \lambda\ell k(\ell)Ke^{2\lambda\theta}}.$$

From now on we need the following assumptions:

(A4) $\mathscr{M}K < h$,

(A5) $K\Big(\dfrac{k_3\ell}{\lambda} + \dfrac{\ell k(\ell)e^{2\lambda\theta}}{e^{\lambda\theta} - 1}\Big) < 1$,

(A6) $\lambda - k_3\ell K - \dfrac{\ln(1 + \ell k(\ell)K)}{\theta} > 0$.

Lemma 6 *Assume that conditions* (A1)–(A2) *and* (A4)–(A6) *are valid. Then almost periodic system (3) has a unique exponentially stable almost periodic solution.*

Proof We denote by $\mathscr{A}P$, $\mathscr{A}P \subset \mathscr{P}C(\mathbb{R}, \mathbb{R}^m)$ the set of all almost periodic functions $\varphi(t)$ satisfying the inequality $||\varphi(t)||_0 \leq h$ with the norm $||\varphi||_0 = \sup_{t\in\mathbb{R}} ||\varphi(t)||$.

Define an operator $\mathscr{E}$ in $\mathscr{A}P$ such that if $\varphi(t) \in \mathscr{A}P$, then

$$\begin{aligned}(\mathscr{E}\varphi)_i(t) = &\int_{-\infty}^{t} X_i(t,s)\left(\sum_{j=1}^{m} b_{ij}(s) f_j(\varphi_j(s)) + c_i(s)\right) ds \\ &+ \sum_{\theta_k<t} X_i(t,\theta_k) J_{ik}(\varphi_i(\theta_k)), \quad i = 1, \ldots, m.\end{aligned}$$

Now, we need to show that $\mathscr{E}(\mathscr{A}P) \subseteq \mathscr{A}P$. If $||\varphi(t)||_0 < h$, then

$$\begin{aligned}||\mathscr{E}\varphi|| &= \sum_{i=1}^{m}\left|\int_{-\infty}^{t} X_i(t,s)\left(\sum_{j=1}^{m} b_{ij}(s) f_j(\varphi_j(s)) + c_i(s)\right) ds + \sum_{\theta_k<t} X_i(t,\theta_k) J_{ik}(\varphi_i(\theta_k))\right| \\ &\leq \sum_{i=1}^{m}\left\{\int_{-\infty}^{t} Ke^{-\lambda(t-s)}\left(\ell k_3|\varphi_i(s)| + k_2\right) ds + \sum_{\theta_k<t} Ke^{-\lambda(t-\theta_k)}\left(\ell k(\ell)|\varphi_i(\theta_k)| + k_4\right)\right\} \\ &\leq \int_{-\infty}^{t} Ke^{-\lambda(t-s)}\left(\ell k_3||\varphi(s)|| + k_2\right) ds + \sum_{\theta_k<t} Ke^{-\lambda(t-\theta_k)}\left(\ell k(\ell)||\varphi(\theta_k)|| + k_4\right) \\ &\leq \mathscr{M}K \leq h.\end{aligned}$$

Next, we shall prove that $\mathscr{E}\varphi$ is almost periodic. By Lemmas 2, 4 and 5, we have

$$\begin{aligned}||\mathscr{E}(\varphi(t+r)) - \mathscr{E}(\varphi(t))|| = &\sum_{i=1}^{m}\left|\int_{-\infty}^{t} X_i(t+r,s+r)\left(\sum_{j=1}^{m} b_{ij}(s+r) f_j(\varphi_j(s+r)) + c_i(s+r)\right) ds\right. \\ &+ \sum_{\theta_k<t} X_i(t+r,\theta_{k+q}) J_{ik+q}(\varphi_i(\theta_{k+q})) \\ &- \int_{-\infty}^{t} X_i(t,s)\left(\sum_{j=1}^{m} b_{ij}(s) f_j(\varphi_j(s)) + c_i(s)\right) ds \\ &\left.- \sum_{\theta_k<t} X_i(t,\theta_k) J_{ik}(\varphi_i(\theta_k))\right| \\ \leq\ &\varepsilon \Gamma_2(\varepsilon),\end{aligned}$$

where $\Gamma_2(\varepsilon)$ is a bounded function of ε. Thus, $\mathscr{E}(\mathscr{A}P) \subseteq \mathscr{A}P$. For arbitrary $\varphi,\ \psi \in \mathscr{A}P$, then

$$
\begin{aligned}
||\mathscr{E}\varphi - \mathscr{E}\psi|| &= \sum_{i=1}^{m} \Bigg| \int_{-\infty}^{t} X_i(t,s) \left(\sum_{j=1}^{m} b_{ij}(s)\Big(f_j(\varphi_j(s)) - f_j(\psi_j(s)) \Big) \right) ds \\
&\quad + \sum_{\theta_k < t} X_i(t,\theta_k)\Big(J_{ik}(\varphi_i(\theta_k)) - J_{ik}(\psi_i(\theta_k)) \Big) \Bigg| \\
&\le \int_{-\infty}^{t} k_3 K \ell e^{-\lambda(t-s)} ||\varphi(s) - \psi(s)|| ds \\
&\quad + \sum_{\theta_k < t} K \ell k(\ell) e^{-\lambda(t-\theta_k)} ||\varphi(\theta_k) - \psi(\theta_k)|| \\
&\le K\Big(\frac{k_3\ell}{\lambda} + \frac{\ell k(\ell) e^{2\lambda\theta}}{e^{\lambda\theta} - 1} \Big) ||\varphi - \psi||_0.
\end{aligned}
$$

By virtue of condition (A5), it follows that $\mathscr{E}$ is contraction. So, there exists a unique almost periodic solution of (3).

Suppose that $z(t) = (z_1, \cdots, z_m)^T$ is an arbitrary solution of (3) and let $z(t) = \varphi(t) - \psi(t) = (\varphi_1 - \psi_1, \cdots, \varphi_m - \psi_m)^T$. Then, by using the integral form of system (3), we have

$$
\begin{aligned}
||z(t)|| \le K e^{-\lambda(t-t_0)} ||z_0|| + \int_{t_0}^{t} k_3 K \ell e^{-\lambda(t-s)} ||z(s)|| ds \\
+ \sum_{\theta_k < t} K \ell k(\ell) e^{-\lambda(t-\theta_k)} ||z(\theta_k)||
\end{aligned}
$$

or

$$
\begin{aligned}
||z(t)|| e^{\lambda(t-t_0)} \le K ||z_0|| + \int_{t_0}^{t} k_3 K \ell e^{\lambda(s-t_0)} ||z(s)|| ds \\
+ \sum_{\theta_k < t} K \ell k(\ell) e^{\lambda(\theta_k - t_0)} ||z(\theta_k)||
\end{aligned}
$$

By the analogue of the Gronwall-Bellman Lemma [54],

$$
||z(t)|| e^{\lambda(t-t_0)} \le K\Big(1 + \ell k(\ell) K \Big)^{i(t_0,t)} e^{k_3 \ell K (t-t_0)} ||z_0||
$$

where $i(t_0, t)$ is the number of points θ_k in $[t_0, t)$. Then, we obtain

$$
||z(t)|| \le K e^{-(\lambda - k_3 \ell K - \frac{\ln(1+\ell k(\ell)K)}{\theta})(t-t_0)} ||z_0||.
$$

Hence, using the condition (A6), the solution of (3) is exponentially stable. □

According to B-equivalence method, the solution $y(t)$ coincides with the solution $x(t)$ of Eq. (1) at discontinuity points $t \in (\theta_k, \theta_{k+1})$. The continuous dependence, in the B-topology, of solutions of Eq. (1) on initial data and the right side implies that $x(t)$ is also almost periodic. Therefore, we have the following result.

Theorem 1 *Assume that conditions (A1)–(A6) are valid. Then an almost periodic system (1) has a unique exponentially stable almost periodic solution.*

3 Robustness Analysis of State-Dependent Impulsive Neural Networks

In this part of the present chapter, we will examine existence, uniqueness and global robust asymptotic stability of the equilibrium point for bidirectional associative memory (BAM) neural networks introduced by Kosko in [33, 34] with state-dependent impulses.

3.1 Model Description and Reduced System

BAM neural networks model consists of two-layers and has many important applications in pattern recognition, signal processing, associative memory, and optimization problems [33, 34]. There are many studies on the global stability and robustness analysis of the BAM neural networks [12–14, 16, 31, 39, 42, 46, 47, 55–58, 65, 66, 74, 78, 79, 82, 84] and references therein. In this part, we analyze existence, uniqueness and global robust asymptotic stability of the equilibrium point for BAM neural networks with state-dependent impulse.

Let $\mathbb{Z}$ and $\mathbb{R}$ be the sets of integers and real numbers. Consider the following impulsive BAM neural networks with variable moments of time:

$$\begin{cases} x_i'(t) = -a_i(t)x_i(t) + \displaystyle\sum_{j=1}^{m} w_{ij}(t) f_j(y_j(t)) + c_i(t) \\ \Delta x_i \mid_{t=\theta_k+\tau_k(x,y)} = d_{ik}x_i + I_{ik}(x), \\ y_j'(t) = -b_j(t)y_j(t) + \displaystyle\sum_{i=1}^{n} h_{ji}(t) g_i(x_i(t)) + d_j(t) \\ \Delta y_j \mid_{t=\theta_k+\tau_k(x,y)} = e_{jk}y_j + J_{jk}(y) \end{cases} \tag{10}$$

where $a_i(t) > 0,\ \ i = 1, 2, \ldots, n$ and $b_j(t) > 0,\ \ j = 1, 2, \ldots, m,\ \ k \in \mathbb{Z},\ \ x \in \mathbb{R}^n,\ \ y \in \mathbb{R}^m, t \in \mathbb{R}$, $\{d_{ik}\}$ and $\left\{e_{jk}\right\}$ are bounded sequences such that $(1 + d_{ik}) \neq 0,\ \ i = 1, 2, \cdots, n$, $(1 + e_{jk}) \neq 0,\ \ j = 1, 2, \cdots, m,\ \ k \in \mathbb{Z}$, $\tau_k(x, y)$ are positive real valued continuous functions defined on $\mathbb{R}^{n+m}$, $k \in \mathbb{Z}$. Moreover, the sequence θ_k satisfies the condition $\theta_k < \theta_{k+1}$, $|\theta_k| \to +\infty$ as $|\mathrm{k}| \to \infty$.

The system (10) is composed of two layers, that is, X-layer and Y-layer, where for $i = 1, 2, \ldots, n$, the $x_i(t)$, denotes the membrane potentials of the set of n neurons in X-layer and for $j = 1, 2, \ldots, m$, the $y_j(t)$ denotes the membrane potentials of the set of m neurons in Y-layer at time t; the continuous, bounded functions $f_j(.)$ and $g_i(.)$ represent the measures of activation to its incoming potentials of the unit j from Y-layer and the unit i from X-layer, respectively, at time t; w_{ij} corresponds to the synaptic connection weight of the unit j on the unit i and h_{ji} corresponds to the synaptic connection weight of the unit i on the unit j; c_i and d_j correspond to the bounded external bias or input from outside the network to the unit i and j, respectively; a_i and b_j denote rate with which the ith unit and jth unit will reset their potentials to the resting state in isolation when it is disconnected from the network and external inputs, respectively. It will be assumed that $a_i, w_{ij}, c_i, b_j, h_{ji}, d_j, I_{ik} : \mathbb{R}^n \to \mathbb{R}^n$, $J_{jk} : \mathbb{R}^m \to \mathbb{R}^m$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, $k \in \mathbb{Z}$ are continuous functions.

In the present study, we do not require smoothness and monotonicity of the activation functions $f_j(.)$ and $g_i(.)$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

We can rewrite the system (10) in the vector state space form as

$$\begin{cases} x'(t) = -A(t)x(t) + W(t)f(y(t)) + C(t) \\ \Delta x\ |_{t=\theta_k+\tau_k(x,y)} = d_k x + I_k(x), \\ y'(t) = -B(t)y(t) + H(t)g(x(t)) + D(t) \\ \Delta y\ |_{t=\theta_k+\tau_k(x,y)} = e_k y + J_k(y) \end{cases} \tag{11}$$

where $x(t) = (x_1(t), x_2(t), \cdots, x_n(t))^T$, $y(t) = (y_1(t), y_2(t), \cdots, y_m(t))^T$,
$A(t) = diag(a_1(t), a_2(t), \cdots, a_n(t))$, $B(t) = diag(b_1(t), b_2(t), \cdots, b_m(t))$,
$W(t) = ((w_{ij}(t))_{n\times m})$, $H(t) = ((h_{ji}(t))_{m\times n})$, $C(t) = (c_1(t), c_2(t), \cdots, c_n(t))^T$,
$D(t) = (d_1(t), d_2(t), \cdots, d_m(t))^T$,
$f(y(t)) = (f_1(y_1(t)), f_2(y_2(t)), \cdots, f_m(y_m(t)))^T$,
$g(x(t)) = (g_1(x_1(t)), g_2(x_2(t)), \cdots, g_n(x_n(t)))^T$,
$\Delta x = (\Delta x_1(t), \Delta x_2(t), ..., \Delta x_n(t))^T$,
$\Delta y = (\Delta y_1(t), \Delta y_2(t), ..., \Delta y_m(t))^T$,
$d_k = diag(d_{1k}, d_{2k}, ..., d_{nk})$, $I_k = (I_{1k}, I_{2k}, ..., I_{nk})^T$,
$e_k = diag(e_{1k}, e_{2k}, ..., e_{mk})$, $J_k = (J_{1k}, J_{2k}, ..., J_{mk})^T$.

From now on, we define a new dependent variable $z(t) = \left[x(t)y(t)\right]$, $z(t) \in \mathbb{R}^{n+m}$, where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$. Then, the system (11) can be written in the form

$$\begin{aligned} &z'(t) = -K(t)z(t) + F(t, z(t)) \\ &\Delta z\ |_{t=\theta_k+\tau_k(z)} = m_k z + S_k(z), \end{aligned} \tag{12}$$

where,

$$K(t) = \begin{bmatrix} A_{n\times n}(t) & 0_{n\times m} \\ 0_{m\times n} & B_{m\times m}(t) \end{bmatrix}, F(t, z(t)) = \begin{bmatrix} W_{n\times m}(t) f_{m\times 1}(y(t)) + C_{n\times 1}(t) \\ H_{m\times n}(t) g_{n\times 1} x((t)) + D_{m\times 1}(t) \end{bmatrix},$$

$$\varDelta z(t) = \begin{bmatrix} \varDelta x(t) \\ \varDelta y(t) \end{bmatrix}, m_k = \begin{bmatrix} (d_k)_{n\times n} & 0_{n\times m} \\ 0_{m\times n} & (e_k)_{m\times m} \end{bmatrix}, S_k(z) = \begin{bmatrix} S_k^1(x) \\ S_k^2(y) \end{bmatrix} = \begin{bmatrix} I_k(x) \\ J_k(y) \end{bmatrix}$$

with $0_{n\times m}, 0_{m\times n}$ are zero matrices. Since the system (12) and the system (10) are equivalent, we can use system (12) to perform our analysis. The following assumptions will be needed throughout the section:

(H1) there exists a Lipschitz constant $\ell > 0$ such that

$$|\tau_k(z_1) - \tau_k(z_2)| + |F(t, z_1) - F(t, z_2)| + |S_k(z_1) - S_k(z_2)| \le \ell |z_1 - z_2|$$

and $|\tau_k(z_1)| \le \ell$ for all $z_1, z_2 \in \mathbb{R}^{n+m}$, $k \in \mathbb{Z}$;

(H2) there exists a positive number $\theta \in \mathbb{R}$ such that $\theta_{k+1} - \theta_k \ge \theta$ holds for all $k \in \mathbb{Z}$ and the surfaces of discontinuity $\varGamma_k : t = \theta_k + \tau_k(z),\ k \in \mathbb{Z}$ satisfy the following conditions:

$$\theta_k + \tau_k(z) < \theta_{k+1} + \tau_{k+1}(z),\ |\theta_k| \to +\infty \quad \text{as} \quad |\mathrm{k}| \to \infty,$$

$$\tau_k((E + m_k)z + S_k(z)) \le \tau_k(z),\ z \in \mathbb{R}^{n+m},$$

where E is an $(n+m) \times (n+m)$ identity matrix;

(H3) $\ell(Nh + M) < 1$, where $N = \sup_{t\in\mathbb{R}} \|K(t)\| < +\infty$, $\|z(t)\| \le h$, $M = \sup_{t\in\mathbb{R}} \|F(t, z(t))\| < +\infty$.

One can make similar discussion which has already been described in Sects. 2.1 and 2.2 as follows: a solution of (12) exists and by virtue of assumptions (H2)–(H3), every solution $z(t),\ ||z(t)|| \le h$ of (12) intersects each surface of discontinuity $\varGamma_k : t = \theta_k + \tau_k(z),\ k \in \mathbb{Z}$, at most once. Furthermore, from the continuation of solutions of ordinary differential equation

$$z'(t) = -K(t)z(t) + F(t, z(t)) \tag{13}$$

and the condition $|\theta_k| \to +\infty$ as $|\mathrm{k}| \to \infty$, one can find that every solution $z(t) = z(t, t_0, z^0),\ (t_0, z^0) \in \mathbb{R} \times \mathbb{R}^{n+m}$, of (12) is continuable on $\mathbb{R}$.

For a fix $k \in \mathbb{Z}$, let $z^0(t) = z(t, \theta_k, z^0)$ be a solution of the system of ordinary differential equations (13). Denote by $t = \xi_k$ the time when the solution of (13) intersects the surface of discontinuity $\varGamma_k : t = \theta_k + \tau_k(z(\xi_k)),\ k \in \mathbb{Z}$. Suppose that $z^1(t) = z(t, \xi_k, (E + m_k)z^0(\theta_k) + S_k(z^0(\xi_k)))$ is also a solution of (13). Next, we define a mapping $U_k(z) : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ such that $U_k(z) = z^1(\theta_k) - (E + m_k)z$ (see also Fig. 1 to clarify the procedure of the contraction of the map U_k.) and construct a system of impulsive differential equations with fixed moments, which has the form

$$\begin{aligned} &v'(t) = -K(t)v(t) + F(t, v(t)) \\ &\Delta v \mid_{t=\theta_k} = m_k v + U_k(v), \end{aligned} \tag{14}$$

where $v(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$, $U_k(v) = \begin{bmatrix} U_k^1(x) \\ U_k^2(y) \end{bmatrix}$, $U_k^1 : \mathbb{R}^n \to \mathbb{R}^n$; $U_k^1 = (U_{1k}^1, U_{2k}^1, \cdots, U_{nk}^1)^T$, $U_k^2 : \mathbb{R}^m \to \mathbb{R}^m$; $U_k^2 = (U_{1k}^2, U_{2k}^2, \cdots, U_{mk}^2)^T$, $k \in \mathbb{Z}$ are continuous functions.

Lemma 7 *Assume that conditions* (H1)–(H3) *are satisfied by (12), then there are mappings $U_k(v) : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^{n+m}$, $k \in \mathbb{Z}$, such that corresponding to each solution $z(t)$ of Eq. (12), there is a solution $v(t)$ of the system (14) satisfying $z(t) = v(t)$ if $t \in \mathbb{R} \setminus \bigcup_{k \in \mathbb{Z}} \widehat{(\xi_k, \theta_k]}$. Moreover, the functions $U_k(y)$ satisfy the inequality*

$$||U_k(z) - U_k(v)|| \le \ell k(\ell)||z - v||,$$

where $k(\ell) = k(\ell, h)$ is a bounded function, uniformly with respect to $k \in \mathbb{Z}$ for all $z, v \in \mathbb{R}^{n+m}$, such that $||z|| \le h$ and $||v|| \le h$.

3.2 Equilibrium for State-Dependent Impulsive BAM Neural Networks

In this section, we will investigate the existence of equilibrium point of the system (14). Now, the system (14) can be written in matrix-vector form as follows:

$$\begin{cases} x'(t) = -A(t)x(t) + W(t)f(y(t)) + C(t) \\ \Delta x \mid_{t=\theta_k} = d_k x + U_k^1(x), \\ y'(t) = -B(t)y(t) + H(t)g(x(t)) + D(t) \\ \Delta y \mid_{t=\theta_k} = e_k y + U_k^2(y) \end{cases} \tag{15}$$

Suppose that $v = [x^*, y^*]^T$ is the equilibrium point of the system (15), that is:

$$\begin{cases} A(t)x^*(t) = W(t)f(y^*(t)) + C(t) \\ \Delta x^* \mid_{t=\theta_k} = d_k x^* + U_k^1(x^*), \\ B(t)y^*(t) = H(t)g(x^*(t)) + D(t) \\ \Delta y^* \mid_{t=\theta_k} = e_k y^* + U_k^2(y^*) \end{cases} \tag{16}$$

It can be concluded that the assumptions on activation functions $g_i(\cdot)$, $f_j(\cdot)$ and conditions $d_k x^* + U_k^1(x^*) = 0$, $e_k y^* + U_k^2(y^*) = 0$ guarantee the existence of an equilibrium point for the system (15) by using the commonly known Brouwer's fixed point theorem. Therefore, one can easily say that $v^* = [x^*, y^*]^T$ is an equilibrium point of the system (15).

Now, we have to show also $v^* = [x^*, y^*]^T$ is an equilibrium point of the system (12).

Denote the set of all zeros of the impulse functions of the system (12) by

$$\Upsilon = \left\{ z \in \mathbb{R}^{n+m} \mid m_k z + S_k(z) = 0,\ k \in \mathbb{Z} \right\}.$$

Here, we need to show $v^* \in \Upsilon$.

Lemma 8 *If $v^* = [x^*, y^*]^T$ is an equilibrium point of the system (15), then v^* is an equilibrium point of the system (12).*

The uniqueness of equilibrium point can be concluded from the global asymptotic stability constructed in the next section.

3.3 Global Robust Asymptotic Stability Analysis

In this part of the chapter, we will obtain sufficient conditions for the system (15) to be robust globally asymptotically stable.

For notational convenience we will shift an equilibrium point $v = [x^*, y^*]$ of the system (15) to the origin by $\overline{x}(t) = x(t) - x^*$, $\overline{y}(t) = y(t) - y^*$. The system (15) can be easily transform into the following form:

$$\begin{cases} \overline{x}'(t) = -A(t)\overline{x}(t) + W(t)\overline{f}(\overline{y}(t)) \\ \Delta\overline{x}\mid_{t=\theta_k} = d_k\overline{x} + \overline{U}^1_k(\overline{x}), \\ \overline{y}'(t) = -B(t)\overline{y}(t) + H(t)\overline{g}(\overline{x}(t)) \\ \Delta\overline{y}\mid_{t=\theta_k} = e_k\overline{y} + \overline{U}^2_k(\overline{y}) \end{cases} \tag{17}$$

where;
$\overline{f}(\overline{y}) = [\overline{f}_1(\overline{y}_1), \overline{f}_2(\overline{y}_2), \cdots, \overline{f}_m(\overline{y}_m)]$ and $\overline{g}(\overline{x}) = [\overline{g}_1(\overline{x}_1), \overline{g}_2(\overline{x}_2), \cdots, \overline{g}_n(\overline{x}_n)]$ where, $\overline{f}_j(y_j(t)) = f_j(\overline{y}_j(t) + y^*_j) - f_j(y^*)$, and $\overline{g}_i(x_i(t)) = g_i(\overline{x}_i(t) + x^*_i) - g_i(x^*)$,
$\overline{U}^1_k(\overline{x}(t)) = [\overline{U}^1_{1k}(\overline{x}_1(t)), \overline{U}^1_{2k}(\overline{x}_2(t)), \cdots, \overline{U}^1_{nk}(\overline{x}_n(t))]$ and
$\overline{U}^2_k(\overline{y}(t)) = [\overline{U}^2_{1k}(\overline{y}_1(t)), \overline{U}^2_{2k}(\overline{y}_2(t)), \cdots, \overline{U}^2_{mk}(\overline{y}_m(t))]$ where,
$\overline{U}^1_{ik}(\overline{x}_i(t)) = U^1_{ik}(x_i(t)) - U^1_{ik}(x^*_i(t))$ and $\overline{U}^2_{jk}(\overline{y}_j(t)) = U^2_{jk}(y_j(t)) - U^2_{jk}(y^*_j(t))$, $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m,\ k \in \mathbb{Z}$, respectively.

Clearly it can be concluded that $\overline{f}_j(0) = 0$ and $\overline{g}_i(0) = 0$ for $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$. In the assumption (A1) we assumed that $|F(t, z_1) - F(t, z_2)| \leq \ell |z_1 - z_2|$ so this implies that there exist positive numbers l^1_j, l^2_i such that $|f_j(\kappa_1) - f_j(\kappa_2)| \leq \ell^1_j|\kappa_1 - \kappa_2|$ and $|g_i(\kappa_1) - g_i(\kappa_2)| \leq \ell^2_i|\kappa_1 - \kappa_2|$, for all $\kappa_1, \kappa_2 \in \mathbb{R}$ $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Also, we assumed the activation functions are bounded that is for any $\kappa_1, \kappa_2 \in \mathbb{R}$ there exist ℓ_j^3, ℓ_i^4 such that $|f_j(\kappa_1)| \le \ell_j^3$ and $|g_i(\kappa_2)| \le \ell_i^4$. Using these properties of activation functions with Euclidean norm $\|\cdot\|$ it follows that:

$\left\|\overline{f}(y)\right\| \le \left\|\Pi^{\sigma^1}\right\| \|y\|$ and $\|\overline{g}(x)\| \le \left\|\Pi^{\sigma^2}\right\| \|x\|$ for all $t \neq \theta_k$, $t \in \mathbb{R}$, where $\Pi^{\sigma^1} = diag(\sigma_1^1, \sigma_2^1, \cdots, \sigma_m^1)$, $\Pi^{\sigma^2} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$.

In order to prove the global asymptotic stability of the equilibrium point of the system (15), it will be enough to prove the global asymptotic stability of origin of the system (17).

As we stated in Sect. 1, there are some deviations in the values of the parameters in the system (15). Since these deviations are bounded in practical experiments, the quantities $a_i(t), b_j(t), w_{ij}(t), h_{ji}(t)$ can be intervalized as follows: For $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m,$

$$\begin{cases} A_I := \left[\underline{A}, \overline{A}\right] = \left\{A = diag(a_i(t)) : 0 < \underline{A} \le A \le \overline{A}, i.e., \underline{a}_i \le a_i(t) \le \overline{a}_i\right\}, \\ B_I := \left[\underline{B}, \overline{B}\right] = \left\{B = diag(b_j(t)) : 0 < \underline{B} \le B \le \overline{B}, i.e., \underline{b}_j \le b_j(t) \le \overline{b}_j\right\}, \\ W_I := \left[\underline{W}, \overline{W}\right] = \left\{W = (w_{ij}(t))_{n\times m} : \underline{W} \le W \le \overline{W}, i.e., \underline{w}_{ij} \le w_{ij}(t) \le \overline{w}_{ij}\right\}, \\ H_I := \left[\underline{H}, \overline{H}\right] = \left\{H = (h_{ji}(t))_{m\times n} : \underline{H} \le H \le \overline{H}, i.e., \underline{h}_{ji} \le h_{ji}(t) \le \overline{h}_{ji}\right\}. \end{cases}$$

Now, we will need the following lemmas and definition to investigate our main results.

Definition 6 The impulsive BAM neural network (15) is called globally robust asymptotic stable if there is a unique equilibrium point $(x^*, y^*) = (x_1^*, x_2^*, \cdots, x_n^*, y_1^*, y_2^*, \cdots, y_m^*)$ and it is globally asymptotically stable for all $A \in A_I, B \in B_I, W \in W_I, H \in H_I$.

In the following, for any real symmetric matrix A, the notation $A > 0$ (respectively, $A \ge 0$) means that A is positive definite (respectively, semi-definite).

Lemma 9 ([82]) *Given any real matrices A, B, C of appropriate dimensions and a scalar $\varepsilon > 0$ such that $0 < C = C^T$.Then the following inequality holds: $A^TB + B^TA \le \varepsilon A^TCA + \varepsilon^{-1}B^TC^{-1}B$ where the superscript T means the transpose of a matrix.*

Lemma 10 ([82]) *(Schur complement) Linear matrix inequality:* $\begin{pmatrix} Q(x) & S(x) \\ S^T(x) & R(x) \end{pmatrix} > 0$ *with,* $Q(x) = Q^T(x)$, $R(x) = R^T(x)$ *is the same as* $R(x) > 0$, $Q(x) - S(x)R^{-1}(x)S^T(x) > 0$.

From now on we need the following assumptions:

(H4) There exist diagonal matrices $\Psi_\alpha = diag(\alpha_1, \alpha_2, \cdots, \alpha_n) > 0, \Psi_\beta = diag(\beta_1, \beta_2, \cdots, \beta_m) > 0$, such that the following linear matrix inequalities holds:

$$\begin{pmatrix} 2\Psi_\alpha \underline{A} - \Pi^{\sigma^2}(H^*)^T(H^*)\Pi^{\sigma^2} & \Psi_\alpha \\ \Psi_\alpha & I \end{pmatrix} > 0$$

and
$$\begin{pmatrix} 2\Psi_\beta\underline{B} - \Pi^{\sigma^1}(W^*)^T(W^*)\Pi^{\sigma^1} & \Psi_\beta \\ \Psi_\beta & I \end{pmatrix} > 0$$

(H5) The impulsive operators satisfy the following conditions

$$d_{ik}\overline{x_i} + \overline{U}^1_{ik}(\overline{x_i}) = -\gamma_{ik}\overline{x_i}(\theta_k), 0 < \gamma_{ik} < 2$$

$$e_{jk}\overline{y_j} + \overline{U}^2_{jk}(\overline{y_j}) = -\delta_{jk}\overline{y_j}(\theta_k), 0 < \delta_{jk} < 2$$

For the sake of convenience, we define $W^* = max\left\{|\underline{W}|, |\overline{W}|\right\}$ and $H^* = max\left\{|\underline{H}|, |\overline{H}|\right\}$.

Lemma 11 *Assume that conditions* (H4) *and* (H5) *are valid. Then, the system (17) is globally robust asymptotically stable.*

Proof We choose the following Lyapunov function:

$$V(t, \overline{x}(t), \overline{y}(t)) = \overline{x}^T(t)\Psi_\alpha\overline{x}(t) + \overline{y}^T(t)\Psi_\beta\overline{y}(t).$$

Along the trajectories of the system (17), evaluating the time derivative of $V(t, \overline{x}(t), \overline{y}(t))$, we have:
When $t \neq \theta_k, k \in \mathbb{Z}$,

$$\begin{aligned} V^+(t) \leq & -\overline{x}^T(t)(A(t)\Psi_\alpha + \Psi_\alpha A(t))\overline{x}(t) - \overline{y}^T(B(t)\Psi_\beta + \Psi_\beta B(t))\overline{y}(t) \\ & + \overline{f}^T(\overline{y}(t))W^T(t)\Psi_\alpha\overline{x}(t) + \overline{x}^T(t)\Psi_\alpha W(t)\overline{f}(\overline{y}(t)) + \overline{g}^T(\overline{x}(t))H^T(t)\Psi_\beta\overline{y}(t) \\ & + \overline{y}^T(t)\Psi_\beta H(t)\overline{g}(\overline{x}(t)). \end{aligned}$$

Using Lemma 9 with $\varepsilon = 1$ and $C = I$ we get

$$\begin{aligned} V^+(t) \leq & -\overline{x}^T(t)(A(t)\Psi_\alpha + \Psi_\alpha A(t))\overline{x}(t) - \overline{y}^T(t)(B(t)\Psi_\beta + \Psi_\beta B(t))\overline{y}(t) \\ & + \overline{f}^T(\overline{y}(t))W^T(t)W(t)\overline{f}(\overline{y}(t)) + \overline{x}^T(t)\Psi_\alpha\Psi_\alpha\overline{x}(t) \\ & + \overline{g}^T(\overline{x}(t))H^T(t)H(t)\overline{g}(\overline{x}(t)) + \overline{y}^T(t)\Psi_\beta\Psi_\beta\overline{y}(t) \\ \leq & \ \overline{x}^T(t)(-A(t)\Psi_\alpha - \Psi_\alpha A(t) + \Pi^{\sigma^2}H^T(t)H(t)\Pi^{\sigma^2} + \Psi_\alpha\Psi_\alpha)\overline{x}(t) \\ & + \overline{y}^T(-B(t)\Psi_\beta - \Psi_\beta B(t) + \Pi^{\sigma^1}W^T(t)W(t)\Pi^{\sigma^1} + \Psi_\beta\Psi_\beta)\overline{y}(t) \\ \leq & \ \overline{x}^T(t)(-2\Psi_\alpha\underline{A} + \Psi_\alpha\Psi_\alpha + \Pi^{\sigma^2}(H^*)^T(t)H^*(t)\Pi^{\sigma^2} +)\overline{x}(t) \\ & + \overline{y}^T(-2\Psi_\beta\underline{B} + \Psi_\beta\Psi_\beta + \Pi^{\sigma^1}(W^*)^T(t)W^*(t)\Pi^{\sigma^1})\overline{y}(t). \end{aligned}$$

Thus using Lemma 10 and $(H4)$ we conclude that $V^+(t) < 0$ for $t \neq \theta_k, k \in \mathbb{Z}$.
When $t = \theta_k, k \in \mathbb{Z}$,

$$\begin{aligned} V(\theta_k+) - V(\theta_k) &= \overline{x}^T(\theta_k+)\Psi_\alpha \overline{x}(\theta_k+) - \overline{x}^T(\theta_k)\Psi_\alpha \overline{x}(\theta_k) + \overline{y}^T(\theta_k+)\Psi_\beta \overline{y}(\theta_k+) \\ &\quad - \overline{y}^T(\theta_k)\Psi_\beta \overline{y}(\theta_k) \\ &= \gamma_k(\gamma_k - 2I)\overline{x}^T(\theta_k)\Psi_\alpha \overline{x}(\theta_k) + \delta_k(\delta_k - 2I)\overline{y}^T(\theta_k)\Psi_\beta \overline{y}(\theta_k). \end{aligned}$$

Using $(H5)$ we get $V(\theta_k+) - V(\theta_k) < 0$, so $V'(t) < 0$.

Hence, we can get $V'(t) < 0$ for $\overline{x}(t) \neq 0$, $\overline{y}(t) \neq 0$ and clearly $V(\cdot)$ is radially unbounded which easily results in that the origin of (17) is globally robust asymptotically stable by standard Lyapunov theorem. □

The same discussion about B-equivalence method which was stated in Sect. 2.3 is valid for our system (10) on $\mathbb{R}^{n+m}$. This implies the following main result.

Theorem 2 *Assume that conditions* (H1)–(H5) *are valid. Then (10) has a unique globally robust asymptotically stable equilibrium point.*

3.4 An Illustrative Example

By means of the B- equivalence method, theoretical results guarantees that simulation of reduced system with fixed moments of impulses is fully adequate to the original system. Now, let us consider the following neural networks system with fixed moments of impulses. In what follows, let $\theta_k = 0.9k + (-1)^k/24,\ k \in \mathbb{Z}$ be the sequence of impulsive action and network parameters defined as

$$\begin{cases} \underline{A} = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.6 \end{pmatrix}, \overline{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1.2 \end{pmatrix}, \underline{W} = \begin{pmatrix} 0.25 & -0.6 \\ -0.16 & 0.09 \end{pmatrix}, \overline{W} = \begin{pmatrix} 0.38 & 0.14 \\ 0.15 & 0.13 \end{pmatrix}, \\ \underline{B} = \begin{pmatrix} 0.63 & 0 \\ 0 & 0.57 \end{pmatrix}, \overline{B} = \begin{pmatrix} 1.3 & 0 \\ 0 & 1 \end{pmatrix}, \underline{H} = \begin{pmatrix} -0.13 & -0.16 \\ 0.07 & 0.08 \end{pmatrix}, \overline{H} = \begin{pmatrix} 0.25 & 0.08 \\ 0.12 & 0.24 \end{pmatrix}. \end{cases}$$

with the activation functions $f_j(y_j(t)) = \frac{1}{5}\tanh(y_j(t))$, $g_i(x_i(t)) = \frac{1}{5}\tanh(x_i(t))$ and $\gamma_{ik} = (\frac{1}{100} + \frac{2}{3i}(\cos(1+k))^2)$, $\delta_{jk} = (\frac{1}{100} + \frac{3}{4j}(\sin(\frac{2}{5}k^3))^2)$, $i, j = 1, 2$. So, we can take $\Pi^{\sigma^1} = \Pi^{\sigma^2} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}$ and $\Psi_\alpha = \Psi_\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Clearly, $0 < \gamma_{ik}, \delta_{jk} < 2(\text{i}, \text{j} = 1, 2.)$ and LMIs

$$\begin{cases} -2\underline{A} + I + \Pi^{\sigma^2}(H^*)^T(H^*)\Pi^{\sigma^2} = \begin{pmatrix} -0.5969 & 0.0028 \\ 0.0028 & -0.1967 \end{pmatrix} < 0, \\ -2\underline{B} + I + \Pi^{\sigma^1}(W^*)^T(W^*)\Pi^{\sigma^1} = \begin{pmatrix} -0.2532 & 0.0030 \\ 0.0030 & -0.1385 \end{pmatrix} < 0. \end{cases}$$

hold. Thus, conditions of Theorem 2 satisfied, so the origin of the specified network is globally robust asymptotically stable. To perform numerical simulation, let us choose A, B, W, H from the indicated intervals above, respectively, and obtain the following system:

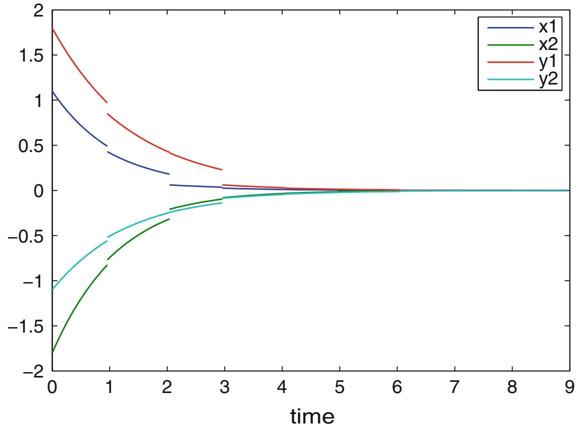


Fig. 2 State trajectory of the system (18). This trajectory is equivalent to the trajectory of the original B-equivalent system with same initial data. As it is seen from the figure, examining the reduced form is more simple and practical

$$\begin{cases} \frac{dx(t)}{dt} = -\begin{pmatrix} 0.9 & 0 \\ 0 & 0.8 \end{pmatrix}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} 0.33 & 0.12 \\ 0.18 & 0.11 \end{pmatrix}\begin{pmatrix} \frac{1}{5}\tanh(y_1(t)) \\ \frac{1}{5}\tanh(y_2(t)) \end{pmatrix}, \ t \neq \theta_k \\ \Delta x(t) = \begin{pmatrix} d_{1k}x_1(\theta_k^-) + U_{1k}^1(x_1(\theta_k^-)) \\ d_{2k}x_2(\theta_k^-) + U_{2k}^1(x_2(\theta_k^-)) \end{pmatrix} = \begin{pmatrix} -(\frac{1}{100} + \frac{2}{3}(\cos(1+k))^2)x_1(\theta_k^-) \\ -(\frac{1}{100} + \frac{1}{3}(\cos(1+k))^2)x_2(\theta_k^-) \end{pmatrix}, \ t = \theta_k, \\ \frac{dy(t)}{dt} = -\begin{pmatrix} 0.65 & 0 \\ 0 & 0.6 \end{pmatrix}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} + \begin{pmatrix} 0.23 & 0.13 \\ 0.9 & 0.16 \end{pmatrix}\begin{pmatrix} \frac{1}{5}\tanh(x_1(t)) \\ \frac{1}{5}\tanh(x_2(t)) \end{pmatrix}, \ t \neq \theta_k \\ \Delta y(t) = \begin{pmatrix} e_{1k}y_1(\theta_k^-) + U_{1k}^2(y_1(\theta_k^-)) \\ e_{2k}y_2(\theta_k^-) + U_{2k}^2(y_2(\theta_k^-)) \end{pmatrix} = \begin{pmatrix} -(\frac{1}{100} + \frac{3}{4}(\sin(\frac{2}{5}k^3))^2)y_1(\theta_k^-) \\ -(\frac{1}{100} + \frac{3}{8}(\sin(\frac{2}{5}k^3))^2)y_2(\theta_k^-) \end{pmatrix}, \ t = \theta_k, \end{cases} \tag{18}$$

By simple calculation, we can see that the LMI conditions of Theorem 1 is satisfied, that is

$$\begin{cases} -2A + I + \Pi^{\sigma^2}(H^*)^T(H^*)\Pi^{\sigma^2} = \begin{pmatrix} -0.7943 & 0.0024 \\ 0.0024 & -0.5989 \end{pmatrix} < 0, \\ -2B + I + \Pi^{\sigma^1}(W^*)^T(W^*)\Pi^{\sigma^1} = \begin{pmatrix} -0.2655 & 0.0070 \\ 0.0070 & -0.1983 \end{pmatrix} < 0. \end{cases}$$

Thus, origin of the system (18) is globally asymptotically stable. This can be seen by simulation in Figs. 2 and 3.

4 Conclusion and Discussion

The problem of investigation of state-dependent impulsive neural networks systems is one of the biggest problem in neural networks theory. In real world problems, the impulses of many systems do not occur at a fixed time, for example, population control systems, saving rates control systems, some circuit control systems [3, 36,

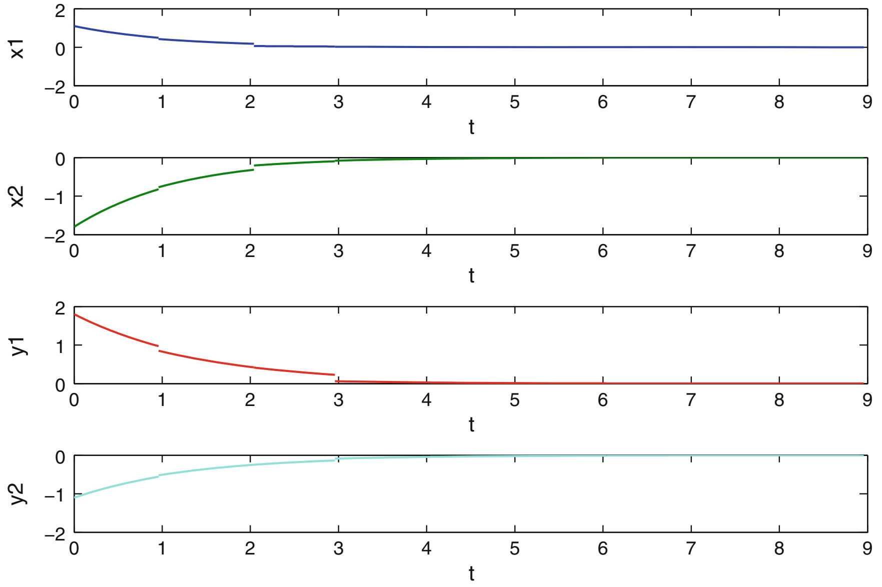


Fig. 3 State trajectories $x_1(t)$, $x_2(t)$, $y_1(t)$, $y_2(t)$ of the system (18). These trajectory is equivalent to the trajectories of the original B-equivalent system with the same initial data

71]. Although, systems with state-dependent impulses commonly exist in both biological and artificial neural networks, it is very difficult to analyze such a system. This is due to the fact that, we know the difference of any two solutions is again a solution, but all of these solutions have different discontinuity time and dynamical behaviors should not be same. So, simple transformations are not allowed for state-dependent impulsive systems. To overcome problem, first we reduced the system to a fix time impulsive system by using B-equivalence method, then we utilize contraction mapping principle, Gronwall-Bellman lemma, some appropriate Lyapunov functions and LMIs. After that, we obtained easily verifiable sufficient conditions for existence and uniqueness of exponentially stable almost periodic solution and robustness of the considered system. Finally, we give one example to show the effectiveness and applicability of our results.

The method developed in this study can be effectively applied to almost all problems in neural networks models such as Cohen-Grossberg neural networks, cellular neural networks, shunting inhibitory cellular neural network, weakly connected neural networks [29]. Since these networks have ability to learn, the method considered in this chapter can be applied to learning theory related to an unsupervised Hebbian-type learning mechanism with/without a forgetting term [20, 24, 26, 35] and several learning algorithms modeled by Amari [11] connected to proposal of Hebb [26].

We note that, in this chapter, we present results of the papers [60, 75], where the first theorems on exponentially stable almost periodic solution and global robust asymptotic stability for state-dependent impulsive neural networks are obtained. Finally, we refer the interested reader to the papers [61, 62] for an extensive treatment on state-dependent impulsive neural networks with time-varying delays.

References

1. Akça, H., Alassar, R., Covachev, V., Covacheva, Z., Al-Zahrani, E.: Continuous-time additive Hopfield-type neural networks with impulses. J. Math. Anal. Appl. **290**, 436–451 (2004)
2. Akhmet, M.U.: Perturbations and Hopf bifurcation of the planar discontinuous dynamical system. Nonlinear Anal.-Theory **60**, 163–178 (2005)
3. Akhmet, M.: Principles of Discontinuous Dynamical Systems. Springer, New York (2010)
4. Akhmet, M.U., Perestyuk, N.A.: The comparison method for differential equations with impulse action. Differ. Equ. **26**, 1079–1086 (1990)
5. Akhmetov, M., Perestyuk, N.: Periodic and almost periodic solutions of strongly nonlinear impulse systems. J. Appl. Math. Mech. **56**, 829–837 (1992)
6. Akhmet, M.U., Yılmaz, E.: Impulsive Hopfield-type neural networks system with piecewise constant argumet. Nonlinear Anal.-Real **11**, 2584–2593 (2010)
7. Akhmet, M.U., Yılmaz, E.: Global exponential stability of neural networks with non-smooth and impact activations. Neural Netw. **34**, 18–27 (2012)
8. Akhmet, M., Yılmaz, E.: Neural Networks with Discontinuous/Impact Activations. Springer, New York (2014)
9. Alzabut, J.O.: Almost periodic solutions for an impulsive delay Nicholson's blowflies model. J. Comput. Appl. Math. **234**, 233–239 (2010)
10. Alzabut, J.O., Stamov, G.Tr., Sermutlu, E.: On almost periodic solutions for an impulsive delay logarithmic population model. Math. Comput. Model. **51**, 625–631 (2010)
11. Amari, S.: Mathematical theory of neural learning. New Gener. Comput. **8**, 281–294 (1991)
12. Arık, S.: An improved robust stability result for uncertain neural networks with multiple time delays. Neural Netw. **54**, 1–10 (2014)
13. Cao, J., Liang, J., Lam, J.: Exponential stability of high-order bidirectional associative memory neural networks with time delays. Phys. D **199**, 425–436 (2004)
14. Cao, J., Ho, D.W.C., Huang, X.: LMI- based criteria for global robust stability of bidirectional associative memory neural networks with time delay. Nonlinear Anal.-Theory **66**, 215–223 (2007)
15. Chen, Z., Ruan, J.: Global stability analysis of impulsive Cohen-Grossberg neural networks with delay. Phys. Lett. A **345**, 101–111 (2005)
16. Chen, A., Huang, L., Liu, Z., Cao, J.: Periodic bidirectional associative memory neural networks with distributed delays. J. Math. Anal. Appl. **317**, 80–102 (2006)
17. Chen, Z., Zhao, D., Ruan, J.: Almost periodic attractor for Cohen-Grossberg neural networks with delay. Phys. Lett. A **373**, 434–440 (2009)
18. Coombes, S., Laing, C.: Delays in activity-based neural networks. Philos. Trans. R. Soc. A **367**, 1117–1129 (2009)
19. Gopalsamy, K.: Stability of artificial neural networks with impulses. Appl. Math. Comput. **154**, 783–813 (2004)
20. Gopalsamy, K.: Learning dynamics in second order networks. Nonlinear Anal.-Real **8**, 688–698 (2007)
21. Gu, H., Jiang, H., Teng, Z.: BAM-type impulsive neural networks with time-varying delays. Nonlinear Anal.-Real **10**, 3059–3072 (2009)

22. Guan, Z.H., Chen, G.: On delayed impulsive Hopfield neural networks. Neural Netw. **12**, 273–280 (1999)
23. Guan, Z.H., Lam, J., Chen, G.: On impulsive autoassociative neural networks. Neural Netw. **13**, 63–69 (2000)
24. Haykin, S.: Neural Networks: A Comprehensive Foundations. Tsinghua Press, Beijing (2001)
25. He, M., Chen, F., Li, Z.: Almost periodic solution of an impulsive differential equation model of plankton allelopathy. Nonlinear Anal.-Real **11**, 2296–2301 (2010)
26. Hebb, D.O.: The Organization of Behaviour. Wiley, NewYork (1949)
27. Ho, D.W.C., Liang, J., Lam, J.: Global exponential stability of impulsive high-order BAM neural networks with time-varying delays. Neural Netw. **19**, 1581–1590 (2006)
28. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sci. Biol. **71**, 2554–2558 (1982)
29. Hoppensteadt, F.C., Izhikevich, E.M.: Weakly Conneceted Neural Networks. Springer, New York (1997)
30. Huang, Z., Luo, X., Yang, Q.: Global asymptotic stability analysis of bidirectional associative memory neural networks with distributed delay and impulse. Chaos Solitons Fractals **34**, 878–885 (2007)
31. Jalab, H.A., Ibrahim, R.W.: Almost-periodic solution for BAM neural networks. Surv. Math. Appl. **4**, 53–63 (2009)
32. Kolmogorov, A.N.: On the Skorohod convergence. Theory Probab. Appl. **1**, 213–222 (1956)
33. Kosko, B.: Bidirectional associative memories. IEEE Trans. Syst. Man Cybern. **18**, 49–60 (1988)
34. Kosko, B.: Adaptive bidirectional associative memories. Appl. Opt. **26**, 4947–4960 (1989)
35. Kosko, B.: Neural Networks and Fuzzy Systems. Prentice-Hall, New Delhi (1992)
36. Lakshmikantham, V., Bainov, D.D., Simeonov, P.S.: Theory of Impulsive Differential Equations. World Scientific, Singapore (1989)
37. Li, X.: Exponential stability of Cohen-Grossberg-type BAM neural networks with time-varying delay via impulsive control. Neurocomputing **73**, 525–530 (2009)
38. Li, Y., Fan, X.: Existence and globally exponential stability of almost periodic solution for Cohen-Grossberg BAM neural networks with variable coefficients. Appl. Math. Model. **33**, 2114–2120 (2009)
39. Li, X., Jia, J.: Global robust stability analysis for BAM neural networks with time-varying delays. Neurocomputing **120**, 499–503 (2013)
40. Li, Y., Yang, C.: Global exponential stability analysis on impulsive BAM neural networks with distributed delays. J. Math. Anal. Appl. **324**, 1125–1139 (2006)
41. Li, K., Zeng, H.: Stability in impulsive Cohen-Grossberg-type BAM neural networks with time-varying delays: a general analysis. Math. Comput. Simul. **80**, 2329–2349 (2010)
42. Li, P., Cao, J., Wang, Z.: Robust impulsive synchronization of coupled delayed neural networks with uncertainties. Phys. A **373**, 261–272 (2007)
43. Li, K., Zhang, L., Zhang, X., Li, Z.: Stability in impulsive Cohen-Grossberg-type BAM neural networks with distributed delays. Appl. Math. Comput. **215**, 3970–3984 (2010)
44. Li, C., Li, C., Liao, X., Huang, T.: Impulsive effects on stability of high-order BAM neural networks with time delays. Neurocomputing **74**, 1541–1550 (2011)
45. Li, G., Yan, Z., Wang, J.: A one-layer recurrent neural network for constrained nonconvex optimization. Neural Netw. **61**, 10–21 (2015)
46. Liu, P., Yi, F., Guo, Q., Wu, W.: Analysis on global exponential robust stability of reaction-diffusion neural networks with S-type distributed delays. Phys. D **237**, 475–485 (2008)
47. Liu, C., Li, C., Liao, X.: Variable-time impulses in BAM neural networks with delay. Neurocomputing **74**, 3286–3295 (2011)
48. Mohamad, S.: Exponential stability in Hopfield-type neural networks with impulses. Chaos Solitons Fractals **32**, 456–467 (2007)
49. Mohamad, S., Gopalsamy, K.: A unified treatment for stability preservation in computer simulations of impulsive BAM networks. Comput. Math. Appl. **55**, 2043–2063 (2008)

50. Mohamad, S., Gopalsamy, K.: Exponential stability preservation in semi-discretisations of BAM networks with nonlinear impulses. Commun. Nonlinear Sci. **14**, 27–50 (2009)
51. Pinto, M., Robledo, G.: Existence and stability of almost periodic solutions in impulsive neural network models. Appl. Math. Comput. **217**, 4167–4177 (2010)
52. Qin, S., Fan, D., Wu, G., Zhao, L.: Neural network for constrained nonsmooth optimization using Tikhonov regularization. Neural Netw. **63**, 272–281 (2015)
53. Samoilenko, A.M., Perestyuk, N.A.: Differential Equations with Impulse Effect. Visca Skola, Kiev (1987) (in Russian)
54. Samoilenko, A.M., Perestyuk, N.A.: Impulsive Differential Equations. World Scientifc, Singapore (1995)
55. Senan, S., Arık, S., Liu, D.: New robust stability results for bidirectional associative memory neural networks with multiple time delays. Appl. Math. Comput. **218**, 11472–11482 (2012)
56. Sheng, L., Yang, H.: Novel global robust exponential stability criterion for uncertain BAM neural networks with time-varying delays. Chaos Solitons Fractals **40**, 2102–2113 (2009)
57. Song, Q., Wang, Z.: An analysis on existence and global exponential stability of periodic solutions for BAM neural networks with time-varying delays. Nonlinear Anal.-Real **8**, 1224–1234 (2007)
58. Song, Y., Han, M., Wei, J.: Stability and Hopf bifurcation analysis on a simplified BAM neural network with delays. Phys. D **200**, 185–204 (2005)
59. Stamov, G.T., Stamova, I.M.: Almost periodic solutions for impulsive neural networks with delay. Appl. Math. Model. **31**, 1263–270 (2007)
60. Şaylı, M., Yılmaz, E.: Global robust asymptotic stability of variable-time impulsive BAM neural networks. Neural Netw. **60**, 67–73 (2014)
61. Şaylı, M., Yılmaz, E.: Periodic solution for state-dependent impulsive shunting inhibitory CNNs with time-varying delays. Neural Netw. **68**, 1–11 (2015)
62. Şaylı, M., Yılmaz, E.: State-dependent impulsive Cohen-Grossberg neural networks with time-varying delays. Neurocomputing **171**, 1375–1386 (2016)
63. Şaylı, M., Yılmaz, E.: Anti-periodic solutions for state-dependent impulsive recurrent neural networks with time-varying and continuously distributed delays. Ann. Oper. Res. (2016). doi:10.1007/s10479-016-2192-6
64. Timofeeva, Y.: Travelling waves in a model of quasi-active dendrites with active spines. Phys. D **239**, 494–503 (2010)
65. Wang, C.: Almost periodic solutions of impulsive BAM neural networks with variable delays on time scales. Commun. Nonlinear Sci. Numer. Simul. **19**, 2828–2842 (2014)
66. Wang, L., Zou, X.: Stability and bifurcation of bidirectional associative memory neural networks with delayed self-feedback. Int. J. Bifurcat. Chaos **15**, 2145–2159 (2005)
67. Xia, Y., Cao, J., Huang, Z.: Existence and exponential stability of almost periodic solution for shunting inhibitory cellular neural networks with impulses. Chaos Solitons Fractals **34**, 1599–1607 (2007)
68. Xia, Y., Cao, J., Lin, M.: New results on the existence and uniqueness of almost periodic solution for BAM neural networks with continuously distributed delays. Chaos Solitons Fractals **31**, 928–936 (2007)
69. Xiang, H., Wang, J., Cao, J.: Almost periodic solution to Cohen-Grossberg-type BAM networks with distributed delays. Neurocomputing **72**, 3751–3759 (2009)
70. Xu, D., Yang, Z.: Impulsive delay differential inequality and stability of neural networks. J. Math. Anal. Appl. **305**, 107–120 (2005)
71. Yang, T.: Impulsive Control Theory. Springer, Berlin (2001)
72. Yang, Y., Cao, J.: Stability and periodicity in delayed cellular neural networks with impulsive effects. Nonlinear Anal.-Real **8**, 362–374 (2007)
73. Yang, F., Zhang, C., Wu, D.: Global stability analysis of impulsive BAM type Cohen-Grossberg neural networks with delay. Appl. Math. Comput. **186**, 932–940 (2007)
74. Yang, D., Liao, X., Hu, C., Wang, Y.: New delay-dependent exponential stability criteria of BAM neural networks with time delays. Math. Comput. Simul. **79**, 1679–1697 (2009)

75. Yılmaz, E.: Almost periodic solutions of impulsive neural networks at non-prescribed moments of time. Neurocomputing **141**, 148–152 (2014)
76. Zhang, X.S.: Neural networks in optimization. Science-Business Media B.V. Springer, New York (2000)
77. Zhang, Y.: Robust exponential stability of uncertain impulsive neural networks with time-varying delays and delayed impulses. Neurocomputing **74**, 3268–3276 (2011)
78. Zhang, Z., Liu, K.: Existence and global exponential stability of a periodic solution to interval general bidirectional associative memory (BAM) neural networks with multiple delays on time scales. Neural Netw. **24**, 427–439 (2011)
79. Zhang, L., Si, L.: Existence and exponential stability of almost periodic solution for BAM neural networks with variable coefficients and delays. Appl. Math. Comput. **194**, 215–223 (2007)
80. Zhang, Y., Sun, J.: Stability of impulsive neural networks with time delays. Phys. Lett. A **348**, 44–50 (2005)
81. Zhou, Q., Wan, L.: Impulsive effects on stability of Cohen-Grossberg-type bidirectional associative memory neural networks with delays. Nonlinear Anal.-Real **10**, 2531–2540 (2009)
82. Zhou, Q., Wan, L.: Global robust asymptotic stability analysis of BAM neural networks with time delay and impulse: an LMI approach. Appl. Math. Comput. **216**, 1538–1545 (2010)
83. Zhang, H., Xia, Y.: Existence and exponential stability of almost periodic solution for Hopfield-type neural networks with impulse. Chaos Solitons Fractals **37**, 1076–082 (2008)
84. Zhang, A., Qiu, J., She, J.: Existence and global exponential stability of periodic solution for high-order discrete-time BAM neural networks. Neural Netw. **50**, 98–109 (2014)

Modelling Native and Invasive Woody Species: A Comparison of ENFA and MaxEnt Applied to the Azorean Forest

Lara Dutra Silva, Hugo Costa, Eduardo Brito de Azevedo, Vasco Medeiros, Mário Alves, Rui Bento Elias and Luís Silva

Abstract Species distribution models are algorithmic tools that relate the distribution and occurrence of a species to the environmental characteristics of the location from where it has been recorded. Those models, also known as ecological niche models, have emerged as an effective tool in spatial ecology, conservation and land management. The Ecological Niche Factor Analysis (ENFA) is one of the common

L.D. Silva (✉) · L. Silva
InBIO, Rede de Investigação em Biodiversidade, Laboratório Associado, CIBIO,
Centro de Investigação em Biodiversidade e Recursos Genéticos, Polo-Açores,
Departamento de Biologia, Universidade dos Açores, 9501-801 Ponta Delgada,
Açores, Portugal
e-mail: laradutrasilva@gmail.com

L. Silva
e-mail: luis.fd.silva@uac.pt

H. Costa
School of Geography, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: lgxhag@nottignham.ac.uk

E.B. de Azevedo
Research Centre for Climate, Meteorology and Global Change (CMMG - CITA-A),
Departamento de Ciências Agrárias, Universidade dos Açores, 9700-042
Angra do Heroísmo, Açores, Portugal
e-mail: eduardo.mv.azevedo@uac.pt

V. Medeiros
Direção Regional dos Recursos Florestais dos Açores, Rua do Contador, 23,
9500-050 Ponta Delgada, Açores, Portugal
e-mail: Vasco.AM.Medeiros@azores.gov.pt

M. Alves
CEO NATURALREASON, Lda, Caminho do Meio Velho, 5-B, 9760-114
Cabo da Praia, Açores, Portugal
e-mail: mario.alves@naturalreason.pt

R.B. Elias
CE3C, Centre for Ecology, Evolution and Environmental Changes/Azorean,
Departamento de Ciências Agrárias, Universidade dos Açores, 9700-042 Angra
do Heroísmo, Açores, Portugal
e-mail: rui.mp.elias@uac.pt

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_20

modelling approaches that are suitable for predicting potential distributions, based on presences only, providing an ecological interpretation based on marginality and specialization. In Maximum Entropy Modelling (MaxEnt) the relative entropy is minimized between the two probability densities defined in the covariate space i.e. estimated from presence data or from landscape. It focuses on relating the environmental conditions of the area where the species is present to the environmental conditions across the area of interest. ENFA has been successfully used in the Azores to model the potential distribution of indigenous and non-indigenous trees. In this paper we use distribution data from one of the most important woody plant invaders in the Azores, *Pittosporum undulatum*, to compare both modelling approaches. We also test both methods when using the selected environmental variables to predict the distribution in other islands and for other species (*Acacia melanoxylon* and *Morella faya*). In general, the two methodologies derived similar predictions. However, our results suggest that the set of environmental variables selected to model the distribution of a species in one particular island will probably have to be adjusted to fit other regions and species.

Keywords Azores · ENFA · MaxEnt · Modelling · SDMs

1 Introduction

In the last two decades, the fascinating question of how plants are distributed in space and time has inspired many ecologists to find convincing explanations. The advances in computational capabilities have provided increasingly greater and more intensive statistical computations than was previously possible [1]. The development and use of numerous statistical techniques in ecology and forestry is extremely important for a better understanding of the species-habitat relationship [2]. This rela tionship is fundamentally tied to the possibility of predicting the potential distribution by relating known species distributions to the spatial distribution of environmental variables [3].

Modelling species habitat requirements has been a particularly helpful tool in the domain of ecosystem management where the identification and protection of areas containing high biological diversity has become crucial, but where species data sets are commonly limited or lacking [4, 5].

Numerous methods can be used to model species distributions [6, 7] that are most often allied to geographic information system (GIS) techniques, allowing the evaluation of their predictive performance [8, 9].

Species distribution models (SDMs) can serve as a tool to ensure consistency in ecological studies, while reducing the time and costs of large-scale studies of biodiversity involving large numbers of species [10]. For example, SDMs have been used to assess the potential distribution of invasive species [11] and to evaluate the possible impact of land use changes on species distribution [12]. In forestry these models were originally designed and used for research purposes, but are presently being developed for use in practical forest management [13, 14]. Robust predictive models of forest biota distribution are important tools to understand ecological theory and the environmental processes affecting species distribution [15].

Human activities have either directly or indirectly influenced almost every part of our world [16]. There are many activities which can affect landscapes in numerous ways, ranging from areas without any significant human impact to urban areas [17], and also vegetation dynamics [18]. Due to globalization, SDMs are being used to predict spatial patterns of biological invasions and prioritize locations for early detection and control of invasion outbreaks [19, 20]. Biological invasions by non-indigenous species are recognized to cause significant losses in the economic value, biodiversity and health of the invaded systems [21, 22]. Globally, invasive species are considered to be one of the most important causes of extinction and decline of indigenous species faced by island ecosystems [23, 24].

In the Azores archipelago, invasive plants can produce dramatic changes in the ecosystems, causing serious problems not only for biodiversity, but also in forestry. These changes are an important threat to biodiversity conservation [25–27], as their effects are emphasized by the vulnerability and peculiarities of the indigenous island biota [28, 29]. No less than 60% of the approximately 1,000 species of vascular plants inhabiting the Azorean islands were introduced by human activities, and are now considered as either naturalized or frequently escaped [30, 31]. Furthermore, the Azorean production forest is dominated by a small number of species and more than 30% of the forested areas are occupied by exotic woodland [32]. In particular, Azorean native vegetation is threatened by several widespread invaders. As a result, usually below 500 m a.s.l., most forest patches are dominated by *Pittosporum undulatum* and *Acacia melanoxylon*, along with few individuals remaining from former natural forests, from species such as *Morella faya* [33]. There is an urgent need to find alternative uses and ways to control alien species [31, 33].

An important step for the construction of species distribution models is the determination of the statistical association between species distribution data and the different independent variables describing the topographic and climatic conditions of the study area, influencing the quality of the final result [34–36]. A variety of statistical algorithms can be used for modelling species potential distributions [37].

The general aim of our research was to use the Azorean islands as models to test the application of SDMs in forestry. The particular aim of this study was to test and compare two different modelling techniques that use presence-only data, the Ecological Niche Factor Analysis (ENFA, [38]) and Maximum Entropy (MaxEnt, [39]), and to identify possible limitations of the modelling tools. We used distribution data from two species that are commonly found in exotic woodlands (*Pittosporum undulatum* and *Acacia melanoxylon*) and one indigenous species (*Morella faya*), in the three largest Azorean islands: Pico, Terceira and São Miguel.

2 Materials and Methods

2.1 *Study Area*

The Azores are a remote and geologically recent archipelago with a total surface area of 2 323 km^2 and consisting of nine volcanic islands, aligned on a west-northwest-east-southeast trend, located in the North Atlantic Ocean at the junction of the Eurasian, African, and North American plates [40]. The archipelago is divided into three groups: the Western Group with Corvo and Flores islands; the Central Group with Faial, Pico, Graciosa, São Jorge, and Terceira islands; and the Eastern Group with São Miguel and Santa Maria islands. The distance between the archipelago and the nearest continent (Europe) is about 1 500 km and the islands span 615 km [32]. The climate is temperate oceanic with a mean annual temperature of 17 °C at sea level. Relative humidity is high and the mean rainfall ranges from 1 500 to more than 3 000 mm per year, increasing with altitude and from east to west [32, 40]. The natural vegetation includes diverse communities, namely coastal vegetation, coastal and inland wetlands, meadows, peat bogs and several types of native forests and scrubs. Human settlement began in the 15th century and since then several activities have altered and affected native plant communities [25, 31].

Pico Island

The island of Pico is located between the coordinates 38° 30' North and 28° 20' West, aged no more than 300 000 years, it is geologically the youngest of the Azores; with a land surface of 449 km^2, it is the largest of the five islands that make up the Central Group [41]. Pico is noted for its eponymous volcano, Ponta do Pico, which is the highest mountain in Portugal with 2 351 m [41, 42]. In geological terms, the tectonic structures of Pico Island are oriented along a west-northwest to east-southeast and a northeast to southwest axis. The main axis controls the main structures, especially the main mountain of Pico, while the secondary axis affect the radial fractures and faults along the central plain and eastern volcano [42, 43].

Terceira Island

Terceira is the third largest island in the Azores archipelago, after São Miguel and Pico. Located at 37° 43' North and 27° 10' West, it is part of the Central Group, being slightly elongated in the east-west direction with an emerged area of about 382 km^2 [44, 45]. The relief increases towards west, where the highest elevation reaches 1 021 m (Serra de Santa Bárbara), and the topography is dominated by four main strato-volcanoes: Cinco Picos, Guilherme Moniz, Pico Alto and Santa Bárbara [44].

São Miguel Island

São Miguel Island is the largest and the most populous of the Azores, located at 37° 50' North and 25° 30' West [46]. Its formation followed a series of volcanic events and it is part of the Eastern Group together with Santa Maria [33]. São Miguel Island extends along the oldest portion (dated about 4 M years old), which starts in the solid mass of Povoação and Nordeste. The highest elevation of the island (Pico da Vara, 1 103 m) is localized in that portion. The island continues towards the west by a series of connected younger volcanic masses, the western most (Sete Cidades) is still active, with the dated from 1652 [33].

2.2 Target Species

Pittosporum undulatum

Pittosporum undulatum Ventenat (Pittosporaceae) (sweet pittosporum, "incenso"), from Australia was introduced in the Azores in the 19th century as an hedge plant in orange tree plantations [32, 47]. It is an invader able to colonize a wide range of habitats, such as tropical and subtropical mountain forest, warm temperate regions of Northern Hemisphere, many islands of the Atlantic and Pacific oceans, and South Africa [48]. It is a shrub or tree up to 15 m tall, and in its native range it is found in different types of *Eucalyptus* forest from Australia, being one of the most abundant species [49]. The canopy is pyramidal, with 3–5 m in diameter and the leaves are evergreen, ovate and with a wavy margin [50]. In the Azores, this species found favourable conditions, invading a considerable portion of the area between sea level and 500 m of altitude, profoundly changing the appearance of the Azorean landscapes [33, 50, 51]. It is one of the species that caused more profound changes in the natural flora, affecting nature reserves, protected landscapes and several native and endemic plant taxa [25]. Also, the endangered Azorean Bullfinch is threatened by *P. undulatum*, and Silva and Tavares [52] showed that only a few introduced arthropod species are able to survive on the invader. The ecological impacts associated with *P. undulatum* invasion should make it one of the priority species for the implementation of control actions in the Azores [32].

Acacia melanoxylon

Acacia melanoxylon R. Br (Fabaceae) (blackwood, "acácia") is an Australian and Tasmanian species that has invaded woodlands and degraded natural habitats [53]. However, it is a valuable cabinet wood species which is commonly found as a canopy or subcanopy tree in a broad range of mixed-species, moist forests on tablelands, and has been spread all over the world because of its ornamental value and the quality of its dark wood [50, 54]. Blackwood was introduced in Southern Europe as an ornamental plant in the 19th century, naturalized in the local habitat and its expansion occurred in the first half of the 20th century through national forestation programs [53–55]. In Europe, the species is considered as invasive, characterized by versatile and highly adaptive tree or root sprouts and with seed germination stimulated by fire [53, 54]. Upon invasion, it establishes quickly in the alien environment, thereby resulting in changes to the structure and dynamics of native ecosystems [53]. It is an evergreen tree with a pyramidal or rounded dense crown, usually reaching

8–15 (20) m high [50]. In the Azores, *A. melanoxylon* has important settlements on Pico Island, in all the northern slopes of the island, and some scattered spots in Terceira and São Miguel islands, however its interest fell sharply with the introduction of *Cryptomeria japonica* in the regional wood market [50].

Morella faya

The origin of *Morella faya* (Aiton) Wilbur (Myricaceae) (fire tree, "faia-da-terra") dates to the Tertiary Mediterranean, what allows considering it as part of a relic flora that once covered Southern Europe and Northern Africa [56]. It is often considered as an Ibero-Macaronesian endemism, being present in the Iberian Peninsula, in the Azores, Madeira and in the Canary Islands. In the end of 1800's, *M. faya* was introduced to Hawaii by Portuguese immigrants [52]. By the 1950's it was considered as a noxious weed, invading rangelands, pasture lands and the natural forests of Hawaii, covering an area of 34 000 ha [52]. *M. faya* is an evergreen shrub or small tree, found on all the Azores islands and is frequently a dominant species in lowland and coastal native vegetation [26, 52]. Nevertheless, it can also be found up to 600 m of altitude and even as high as 1 000 m in Pico Island, while in the Canary Islands and Madeira, it is found up to 1 300 m [47, 52]. It occurs in exposed habitats on various substrates (siliceous soils), on cliffs near the coast or on lava flows and in humid environments [52]. In the Azores, this species can reach up to 20 m high (diameter exceeding the 35 cm) [50]. The plant has ecological and morphological peculiarities which enable a beneficial role of this native species in the Azores. It is useful for cloud water infiltration, soil nutrient fixation and as a source of food and shelter for birds and endemic insects [56]. However, *M. faya*, distribution in the Azores is being reduced as a consequence of human activities (changes in land cover) and by the spread of *P. undulatum* [31, 56, 57]. According to the Forest Services [58], *M. faya* is currently present in a total of 22% of the forested area in the Azores archipelago but it is the dominant species only in 5% of that area [26]. For this reason it is fundamental to preserve *M. faya* and the associated flora and fauna [26, 56].

2.3 Ecogeographical Variables

The environmental variables needed for a SDM must have the potential to ecologically explain the distribution of the species. These can be classified into topographic, climatic or geologic, defining the environmental factors that delimit the conditions favourable for the species being present at a given location [59, 60]. Eighty ecogeographical variables (EGV) (topography, climate and land use) were obtained from different sources and used to describe 100 × 100 m cells covering the entire study area.

Topographical variables were derived based on a digital elevation model (DEM) available in the CIELO model [61–63] (see http://www.climaat.angra.uac.pt). The variables used were digital elevation model, aspect, slope, curvature, flow accumulation, summer hill shade and winter hill shade. Curvature is the second derivative of the surface, which highlights the local geomorphology, specifically the flatness, convexity and concavity of the surface. Flow accumulation is a simulation of the superficial flow by considering that all raster cells flow into each downslope cell in the DEM. Hill shade is a simulation of the lighting conditions on the surface dictated by the topography and the position of the Sun (the Summer and Winter solstices were considered).

Climatic variables were derived from CIELO model developed by Azevedo [63]. This model simulates the climatic variables in an island reproducing the thermodynamic transformations experienced by an air mass crossing the orography, and simulates the evolution of the air parcel's properties starting from the sea level [61]. The model consists of two main sub-models. One, relative to the advective component simulation, assumes the Foehn effect to reproduce the dynamic and thermodynamic processes. The second concerns the radiative component and energy balance as affected by the orographic clouds and by topography. The climatic variables selected, were air temperature, precipitation, relative humidity with maximum and minimum extremes, seasonal variation, annual and monthly averages and a summary obtained by Principal Component Analysis (PCA). Land use variables included [26, 27]: (i) one variable expressing the distance to six types of land use (forest; natural vegetation; pastureland; agriculture; barren/bare areas; and urban/industrial areas); and (ii) an ordinal variable that increased from more artificial land uses to more forest-like land uses. The files containing EGV information were converted into raster-based IDRISI format [64] and into ASCII raster file for further analyses in software Biomapper version 4.0 [65] and software MaxEnt version 3.3.3, respectively [39]. Then a preliminary analysis was performed to investigate which variables would be more suitable for modelling. There are several reasons to reduce the number of variables in a model, such as: to minimize computing time, to minimize the amount of correlated variables that can cause overfitting, and to increase transferability of the models (e.g. for different regions). Since distribution models are correlative, it may be possible to allow the creation of more generalized responses to the environment with a balance between underfitted models with few parameters and overfitted models with too many correlates [66]. Furthermore, the number of explanatory factors should be reduced to a limited set of more significant and less correlated variables to avoid the decrease of transferability effectiveness [67]. Finally, the reduction of correlated variables allows to better understand the causal relationships of the models [9, 67, 68]. Thus, the preliminary analysis of the available EGV has led to the selection of 28 EGV (Table 1) used to proceed with modelling based on topography, climate and land use. This was largely based on correlation and principal component analysis, as well as on preliminary testing of more than 400 possible EGV combinations, using Biomapper.

The models were tested in a sequential manner, starting with topographic or climatic EGV only, combining the selected EGV, and finally adding the land use EGV [69]. During the three phases, the various models arising from different combinations of EGV were evaluated using the methods available in Biomapper and MaxEnt, which are the continuous Boyce curve [34] and the Area Under the Curve (AUC) obtained for the Receiver Operating Characteristic (ROC) curve [39].

Table 1 List of the 28 EGV used to model indigenous and non-indigenous woody species

Variable category	Variables	Code	Unit
Topographical	Digital elevation model	DEM	m
	Aspect	ASP	°
	Slope	SLP	%
	Curvature	CRV	
	Flow accumulation	FLA	
	Summer hill shade	SHS	
	Winter hill shade	WHS	
Climatic	Annual minimum temperature	TMIN	°C
	Annual mean temperature	TM	
	Annual maximum temperature	TMAX	
	Annual temperature range	TRA	
	Annual mean temperature range	TMRA	
	Annual minimum relative humidity	RHMIN	%
	Annual mean relative humidity	RHM	
	Annual maximum relative humidity	RHMAX	
	Annual relative humidity range	RHRA	
	Annual mininum precipitation	PMIN	mm
	Annual mean precipitation	PM	
	Annual maximum precipitation	PMAX	
	Annual precipitation range	PRA	
	Annual mean precipitation range	PMRA	
Land use	Distance to forest	DL 1	
	Distance to natural vegetation	DL 2	
	Distance to pastureland	DL 3	
	Distance to agriculture	DL 4	
	Distance to barren/bare areas	DL 5	
	Distance to urban/industrial areas	DL 6	
	An ordinal variable that has the highest score for habitats of forest type and lowest for urban/social areas	ODL	

2.4 Species Data

Presence-only data sets are frequently the only type of available data containing species occurrence information [70]. The frequent occurrence of presence-only data is due in part to time or financial constraints as well as to data collection strategies aimed at inventories instead of statistical analysis [71]. In our study, we only consider presences, since absences might not reflect unfavourable environment but be a consequence of land cover changes (i.e. removal of native species, as *M. faya*); or the species might not yet have reached equilibrium with the environment (i.e.

Table 2 Number of species records used in modelling, per island and species. PU: *Pittosporum undulatum*; AM: *Acacia melanoxylon*; MF: *Morella faya*

	Pico			Terceira			São Miguel		
	PU	*AM*	*MF*	*PU*	*AM*	*MF*	*PU*	*AM*	*MF*
Entire dataset	12 922	1 422	6 701	1 885	433	402	5 668	4 867	666
Dataset (75%)	9 692	1 067	5 026	1 414	325	302	4 251	3 650	500
Testing sample	3 231	356	1 675	417	108	101	1 417	1 217	167
Validation sample	2 423	267	1 256	353	81	75	1 063	913	125
Training sample (100%)	7 269	800	3 769	1 060	244	226	3 188	2 738	375
Training sample (75%)	5 451	600	2 827	795	183	170	2 391	2 053	281
Training sample (50%)	3 634	400	1 885	530	122	113	1 594	1 369	187
Training sample (25%)	1 817	200	942	265	61	57	797	684	94
Training sample (12.5%)	909	100	471	133	30	28	399	342	47

invasive species such as *P. undulatum*). Based on distribution data from the Forest Services [58], we obtained the points of occurrence of the target species, by using all the polygons identified in the Forest Inventory. The data, provided in vector format (shapefile), were converted to raster (RST) and CSV formats for further analysis. All biological data was geocoded to a grid of 100 × 100 m in order to match the EGV used.

We then proceeded to a random division of the data into two sets, each containing 25 and 75% of the entire data. The former was used as validation data set and the latter was further divided into testing and training sets (25 and 75% respectively). Finally, the training set was randomly and progressively reduced to 75, 50, 25 and 12.5% of its size to test the reliability of the models using limited training data (Table 2, Fig. 1). A completely random selection of samples was conducted in software QGIS Valmiera 2.2, using an appropriate vector function.

2.5 Modelling

Ecological Niche Factor Analysis (ENFA)

ENFA is a presence-only multifactor analysis, which compares, in the multidimensional space of ecological variables, species distribution to a reference set describing the whole study area [38]. The transformation of EGV into a set of uncorrelated factorial axes allows the introduction of the ecological concepts of marginality and

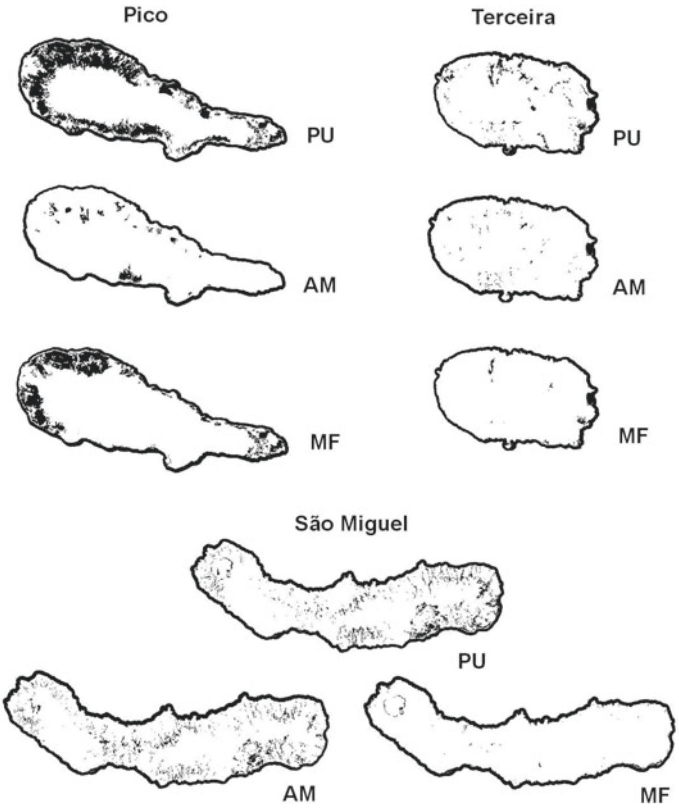


Fig. 1 Distribution data of *P. undulatum* (PU), *A. melanoxylon* (AM) and *M. faya* (MF) in three Azores islands

specialization, two types of factors extracted with biological significance. The first factor extracted gives the marginality coefficient, describing the standardized difference between the average conditions at species sites and in the entire study area. Marginality ranges from -1 to $+1$ and indicates the rarity of the conditions selected by the target species within the study area. Positive values show a species' optimum to be higher than the average conditions in the study area. The second factor, specialization, (varying generally from 0 to ∞) compares the variance of the environment in areas where the species is present with the global variance in the study area. Successive factors explain the remaining specialization in decreasing amounts. A high value of specialization indicates a narrow niche breadth in comparison with the available conditions [34, 38, 72]. Finally, a Habitat Suitability (HS) map with values ranging from 0 to 100 is built, comparing the position of each cell in the study area to the distribution of presence cells on the different factorial axes [38, 65]. We compared three algorithms: median, distance geometric mean and harmonic mean. The median algorithm makes the assumption that the best habitat is at the median of the species distribution on each factor and that these distributions are symmetrical [20, 34]. The distanced geometric mean algorithm and harmonic mean make no assumption about the shape of the species distribution [34].

The software Biomapper version 4.0 [65] (see http://www.unil.ch/biomapper) was used to perform the modelling. Only the principal factors generated in ENFA are relevant to calculate HS values, and hence these were determined based on their eigenvalues compared to the broken-stick distribution [73]. The principal factors retained ranged between 2 and 4 across the various models produced.

Maximum Entropy (MaxEnt)

MaxEnt is a machine learning model that uses presence-only data to predict species distributions based on the principle of maximum entropy [74, 75]. The prediction of the model indicates the areas within the study region that satisfy the requirements of the species ecological niche [76]. We utilized the MaxEnt software version 3.3.3 [39] (see http://www.cs.princeton.edu/schapire/maxent) using a maximum of 500 iterations, autofeatures and the cumulative output [39, 76]. For each species, 20 models (replicates) were generated via bootstrapping [3, 76]. We used 75% of records for training and 25% for testing. The ROC (Receiver Operating Characteristic) curve proposed by Peterson et al. [77] was used to assess the predictive performance of the models. In addition, the continuous Boyce index was also calculated to provide a means of comparison between the results of the ENFA and MaxEnt.

In order to determine the importance of each environmental variable, the Jackknife method available in the software was used [39].

Model evaluation

Verification and validation are essential in the construction SDMs, avoiding the risk of accepting as “true” models that have gross errors [78]. To assess the robustness and predictive ability of the models, the method proposed by Boyce et al. [79] and improved by Hirzel et al. [34] was used. The validation criteria considered was based on the analysis of continuous Boyce indexes and curve shape: (i) the standard deviation around the curves should be narrow; (ii) the curves should be linear and ascending; and (iii) the O (Observed)/ E (Expected) ratio should be high [26, 34]. Ideally, a good HS model produces a monotonically increasing curve and its goodness of fit is measured by the Spearman rank correlation coefficient [26, 34].

MaxEnt results for each model were also evaluated by using the ROC curve [39, 77, 80] although this method is not free from errors [81, 82]. A good model is defined by a curve that maximizes sensitivity for low values of the false - positive fraction [83]. Model quality is quantified by the AUC, with values close to 0.5 indicating a classifier no better than random expectation a values close to 1 indicating good discriminating power [84].

McNemar’s test was used to check whether the predictions of the two modelling techniques coincided. This non-parametric test assesses if a statistically significant change in proportions have occurred on a dichotomous trait at two points on the same population. It is built on the off diagonal elements of 2 × 2 confusion matrices and based on a χ^2 test with one degree of freedom. For that, we analysed 2 × 2 confusion matrices discriminating the two modelling approaches (ENFA and MaxEnt) and the number of cells with HS values above or below the third quartile. Habitat Suitability (HS) was expressed in a continuous raster map of HS scores ranging from 0 to 100.

We first tested the EGV sets with occurrence data for *P. undulatum* and Pico Island, and afterwards tested both modelling approaches, ENFA and MaxEnt, by using the same EGV sets with occurrence data from Terceira and São Miguel islands, and also with occurrence data for *A. melanoxylon* and *M. faya* for the same three islands.

3 Results

3.1 Selected EGV

When developing the models for *P. undulatum* and Pico Island, according to the continuous Boyce index and to the AUC, the models of higher predictive ability were obtained by using the EGV sets described in Table 3.

3.2 Ecological Niche Factor Analysis (ENFA)

Regarding *P. undulatum* in Pico Island, the three algorithms available for habitat suitability computation in Biomapper, namely median, geometric mean and harmonic mean all gave similar results regardless of the EGV set or the sample size (results not

Table 3 Sets of EGV that produced models with highest predictive ability

Variable category	Variables	Code	EGV set			
			1	2	3	4
Topographical	Digital elevation model	DEM	+		+	+
	Slope	SLP	+		+	+
	Curvature	CRV	+		+	+
	Flow accumulation	FLA	+		+	+
	Summer hill shade	SHS	+		+	+
	Winter hill shade	WHS	+		+	+
Climatic	*Temperature*					
	Annual mean temperature	TM		+	+	+
	Annual temperature range	TRA		+	+	+
	Humidity					
	Annual mean relative humidity	RHM		+	+	+
	Annual relative humidity range	RHRA		+	+	+
	Precipitation					
	Annual mean precipitation	PM		+	+	+
	Annual precipitation range	PRA		+	+	+
Land use	Distance to natural vegetation	DL 2				+

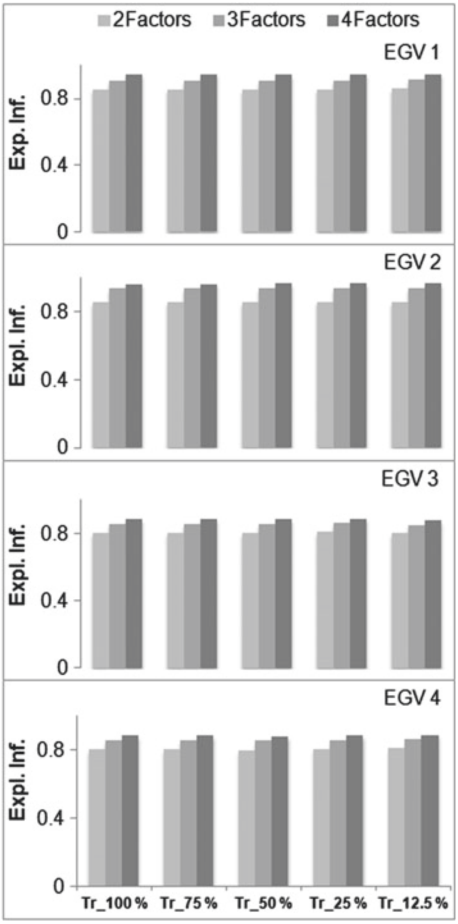


Fig. 2 Modelling results for *Pittosporum undulatum* in Pico Island using Biomapper. Total information explained (Expl. Inf.) of the original EGV set used depending on the factors selected and the size of the training (Tr) data sets

shown). We thus selected the median algorithm for presenting the results, showing that the amount of explained information was generally high and increased with the number of factors included (Fig. 2).

Global marginality values were lower than 1 for *P. undulatum*. The values ranged between 0.387 for EGV set 1 and 0.875 for EGV set 4. The values of specialization ranged between 1.599 for EGV set 1 and 1.654 for EGV set 4 (Table 4).

3.3 Maximum Entropy (MaxEnt)

With regard to the results for *P. undulatum* in Pico Island, small differences in AUC were observed between the four EGV sets and between training samples of decreasing

Table 4 Results of the ENFA applied to *Pittosporum undulatum* in Pico Island, according to the EGV set. Marginality, specialization and tolerance for each EGV set

EGV set	Marginality	Specialization	Tolerance (1/Specialization)
1	0.387	1.599	0.626
2	0.774	1.684	0.594
3	0.865	1.695	0.590
4	0.875	1.654	0.604

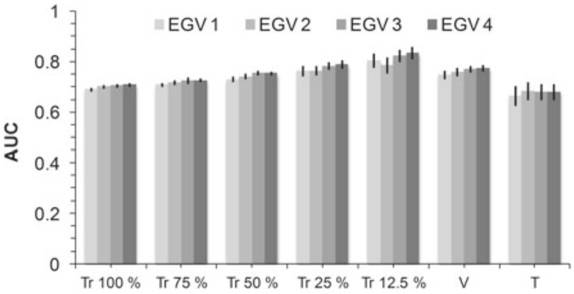


Fig. 3 Modelling results for *Pittosporum undulatum* in Pico Island using MaxEnt. Area Under the Curve (AUC), depending on the EGV set used, and the size of the training (Tr), validation (V) and test (T) data sets. Error bars corresponds to ± 4 sd (Standard deviation)

size. The results obtained when using training, validation and test data sets were similar (Fig. 3).

Indeed the AUC values generally increased with the addition of environmental variables (EGV set 1 to 4) and sampling reduction (training set of 100% to training set of 12.5%). While the differences in AUC values were very small, the changes may still be meaningful ecologically. In the case of the EGV set 4, the performance of MaxEnt was better across the entire spectrum of the EGV included, when compared to the models derived using the other EGV sets. This information was given by the analysis of variable contributions and the results of the Jackknife test of variable importance (Fig. 4).

3.4 Comparative Performance of ENFA and MaxEnt

The models derived by using EGV set 4 were among the best, according to the AUC and the Boyce index (value and curve shape). Figure 5 allows to compare the monotonic curve obtained when using EGV set 4, indicating a good predictive model, with a sawtooth curve associated to a low quality model. Therefore, we analysed the contribution of each EGV included in EGV set 4.

Table 5 explains how the extracted factors in the ENFA are correlated with the EGV. The first factor is highly and negatively correlated with relative humidity and precipitation, and highly and positively correlated with temperature. The digital elevation model is negatively correlated with the marginality factor and is highly corre-

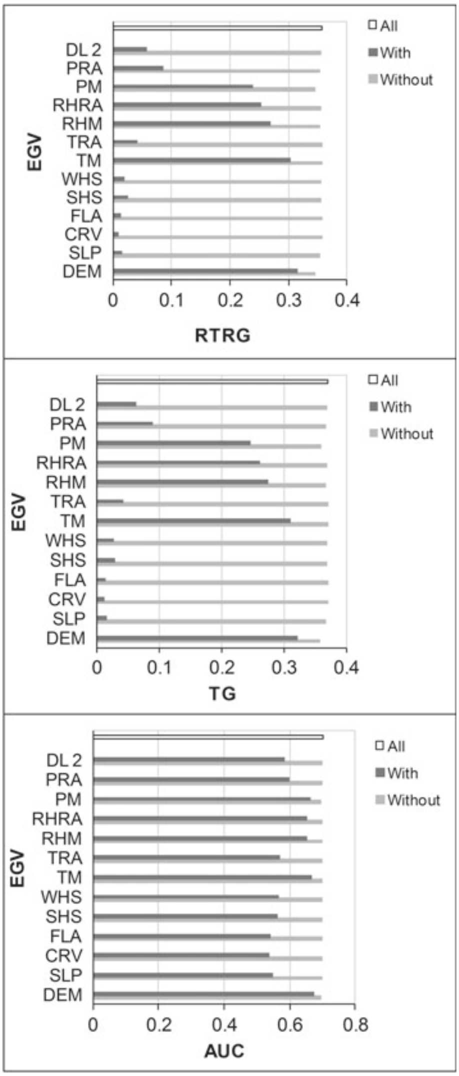


Fig. 4 Analysis of variable contribution for model EGV set 4 using MaxEnt for *Pittosporum undulatum* in Pico Island (Tr_100%). Jackknife of regularized training gain (RTRG) providing variable importance when used in isolation (*top*); Jackknife of test gain (TG), using test instead of training data (*middle*); Jackknife of AUC (Area Under the Curve) providing the contribution of each variable for AUC (*bottom*). All: with all the variables; With: with the variable only; Without: without the variable

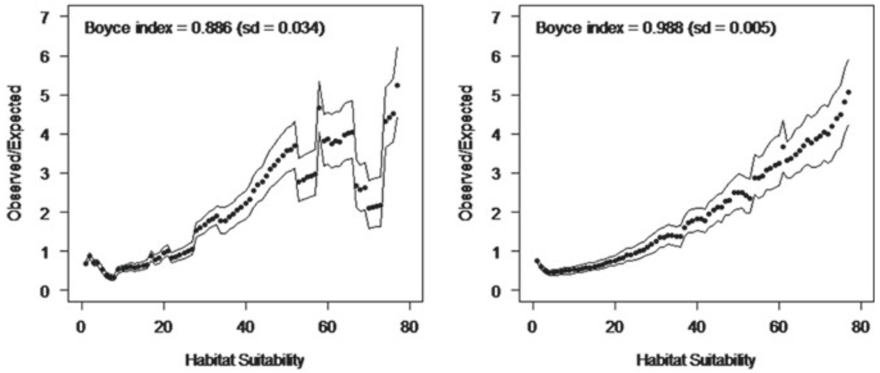


Fig. 5 Results of continuous Boyce curves. HS models for EGV set 1 (2 Factors), sawtooth curve (*left*) and EGV set 4 (4 Factors), monotonic curve (*right*). *Solid black dots* mean continuous Boyce curves; *Black lines* 95% confidence intervals; sd: Standard deviation

lated with the specialization factors (Table 5). These results suggest that *P. undulatum* avoids the high mountains, where temperatures are low and relative humidity and rainfall are high.

Regarding MaxEnt, the contribution of each EGV to the predicted models was strongly dependent on the size of training data set, but digital elevation model (DEM), annual mean temperature (TM) and annual mean relative humidity (RHM) generally contributed most to the final model (Table 5).

Similarly to what was found for the ENFA, when using MaxEnt *P. undulatum* was shown to prefer lower elevations, relatively high mean annual temperatures, and relatively low mean relative humidity, according to the response curves given by MaxEnt (Fig. 6).

We compared ENFA and MaxEnt results using the continuous Boyce index since AUC might present some drawbacks [85]. The comparison showed similar results for both approaches when using training, validation and test data sets (Fig. 7).

However, there was a tendency for the ENFA (Boyce index 0.768-0.988) to perform slightly better than MaxEnt (Boyce index 0.598-0.977). There was a general increase of the Boyce index with the addition of more EGVs for both modelling approaches. Reducing the sample size of the training data set lead to a reduction in the Boyce index. For MaxEnt the Boyce index showed a larger standard deviation than for ENFA, implying lower reliability, and the models with the largest index did not coincide for both techniques (i.e., EGV set 4 for ENFA and EGV set 2 for MaxEnt). When comparing the two approaches with McNemar's test, the differences sharply decreased as more factors were included in ENFA (Table 6).

Table 5 Scores of the EGV on the first four axes of the Ecological Niche Factor Analysis and the estimates of the contributions and permutation importance of each EGV for the models derived using MaxEnt for *Pittosporum undulatum* in Pico Island (EGV set 4). Factor 1 (F1) includes 100% marginality and part of the specialization. Factors 2-4 (F2-F4) explain species specialization in decreasing order

Variable category	Variables	Code	ENFA				MaxEnt	
			Score matrix					
			F1	F2	F3	F4	% contribution	% permutation importance
Topographical	Digital elevation model	DEM	−0.412	0.677	0.437	0.528	57.90	27.90
	Slope	SLP	−0.099	0.013	0.168	0.220	0.80	6.10
	Curvature	CRV	−0.034	0.016	−0.008	0.088	0.20	0.70
	Flow accumulation	FLA	0.111	0.012	−0.011	−0.003	0.10	0.50
	Summer hill shade	SHS	0.046	−0.044	−0.045	0.725	0.20	3.00
	Winter hill shade	WHS	−0.005	0.036	0.010	−0.136	0.20	1.90
Climatic	Annual mean temperature	TM	0.419	−0.351	−0.138	0.145	17.20	1.80
	Annual temperature range	TRA	−0.229	−0.204	−0.020	−0.221	0.00	0.30
	Annual mean relative humidity	RHM	−0.404	−0.397	−0.264	−0.002	14.70	16.10
	Annual relative humidity range	RHRA	0.391	0.455	−0.292	0.205	1.20	4.90
	Annual mean precipitation	PM	−0.405	−0.089	−0.758	−0.044	6.00	29.20
	Annual precipitation range	PRA	0.271	0.015	−0.171	−0.045	1.30	6.00
Land use	Distance to natural vegetation	DL 2	0.150	0.023	0.004	0.068	0.20	1.50

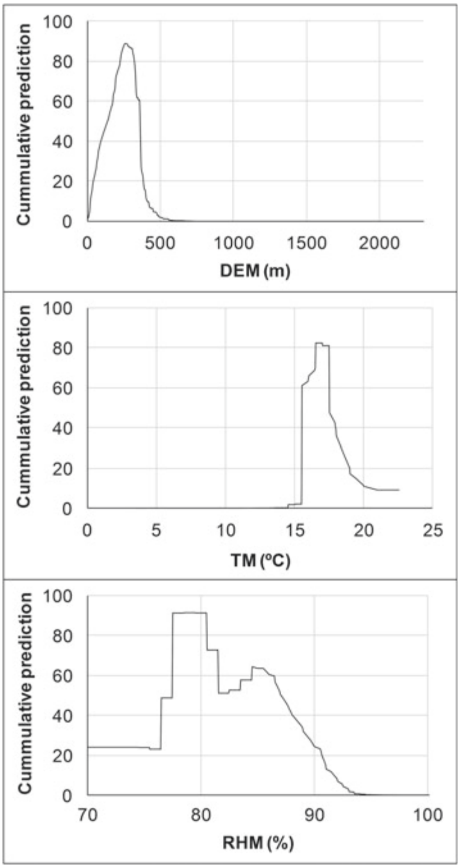


Fig. 6 Response curves of *Pittosporum undulatum* to the three EGV considered as more important in Pico Island (EGV set 4; Tr_100%) obtained using MaxEnt. DEM, Digital elevation model (*top*); TM, Annual mean temperature (*middle*); RHM, Annual mean relative humidity (*bottom*)

3.5 Testing the EGV Sets for Other Islands/species

Figures 8, 9 and 10 show the predicted habitat suitability maps corresponding to the models built using EGV set 4. In general, both modelling approaches originated similar potential distribution maps. However, MaxEnt predictions seemed to loose more quality than ENFA when applying the same EGV sets to the new islands/species, since although the value of the AUC was not considerably affected, the value of the continuous Boyce index declined in some cases (Table 7).

4 Discussion

In previous studies in the Azores, using ENFA and Generalized Linear Models (GLMs), it was reported that the current distribution of *P. undulatum* in São Miguel

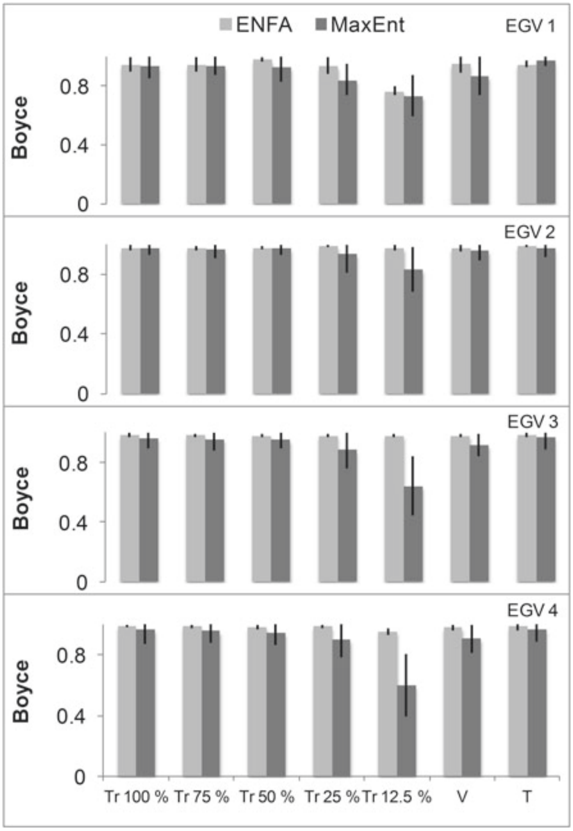


Fig. 7 Modelling results for *Pittosporum undulatum* in Pico Island using ENFA (4 Factors) and MaxEnt. Continuous Boyce index, according to the EGV set used, and the size of the training (Tr), validation (V) and test (T) data sets. Error bars corresponds to ± 2 sd (Standard deviation)

Table 6 Results of a McNemar's test comparing the predicted habitat suitability maps for *Pittosporum undulatum*, derived using ENFA and MaxEnt. The obtained chi-square distance is shown, for the four EGV sets tested, and for the ENFA models derived by including different numbers of factors, and different percentages of the training data set (25 or 75%). The test was used to evaluate the degree of coincidence of both methods for indicating the cells with habitat suitability above or below the third quartile. Low values indicate similarity between the ENFA and MaxEnt results

EGV set	ENFA *versus* MaxEnt					
	2 Factors		3 Factors		4 Factors	
	χ^2(25%)	χ^2 (75%)	χ^2 (25%)	χ^2 (75%)	χ^2 (25%)	χ^2 (75%)
1	272.1	677.6	152.7	38.2	4.5	13.9
2	111.3	80.9	9.6	3.2	1.5	3.9
3	160.4	191.5	26.2	11.4	4.8	1.2
4	238.1	125.5	73.2	27.4	1.6	28.7

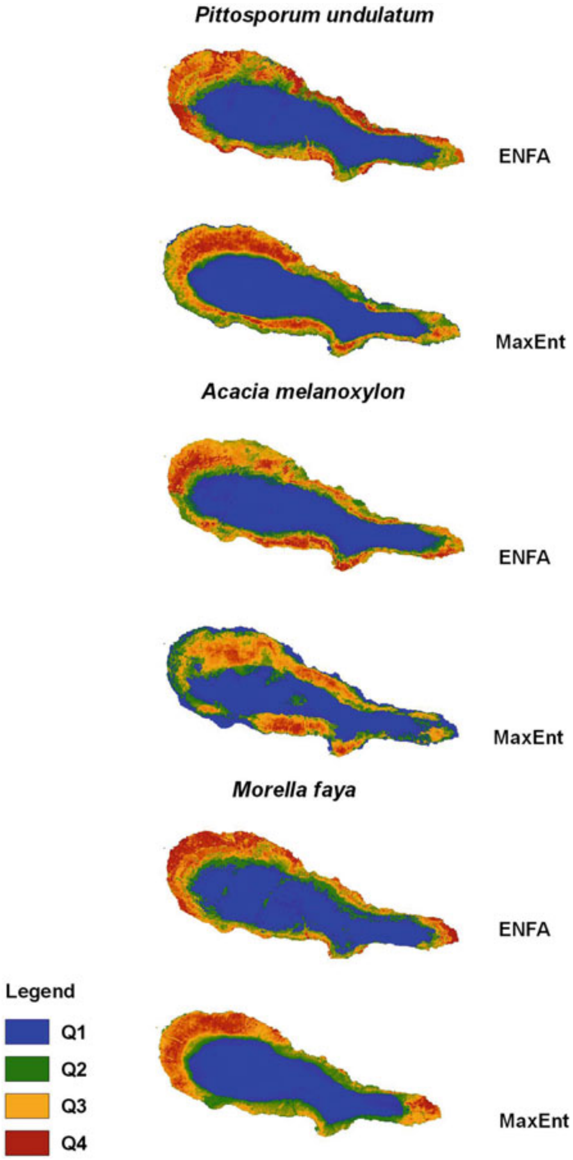


Fig. 8 Habitat suitability maps for *P. undulatum*, *A. melanoxylon* and *M. faya* based on the EGV set 4, derived with occurrence data from Pico Island, and applying two modelling approaches, ENFA and MaxEnt. Habitat suitability was reclassified into four levels, according to the respective quartiles

Island is related to climate, altitude and some human activity effects. Analysis of the areas under risk of invasion showed that protected areas are under potential threat.

Costa et al. [26] used ENFA to evaluate whether and where areas currently occupied by *P. undulatum* could also be favourable habitat for *M. faya*, thus providing support for future management actions. The two species were shown to have quite similar environmental preferences, which correspond mainly to coastal and lowland

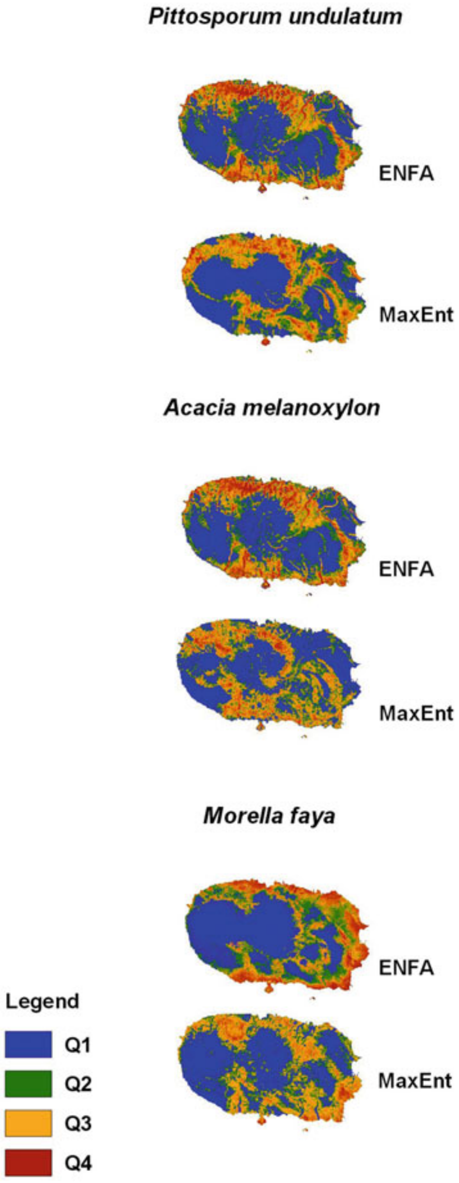


Fig. 9 Habitat suitability maps for *P. undulatum*, *A. melanoxylon* and *M. faya* based on the EGV set 4, derived with occurrence data from Terceira Island, and applying two modelling approaches, ENFA and MaxEnt. Habitat suitability was reclassified into four levels, according to the respective quartiles

areas and forested habitats characterized by relatively high temperature and relatively low but widely variable relative humidity values. The same authors [27] used the ENFA with presence-only data sets of two top invasive woody species (*P. undulatum* and *A. melanoxylon*), showing ENFA to be a suitable method for modelling environmental weed distributions.

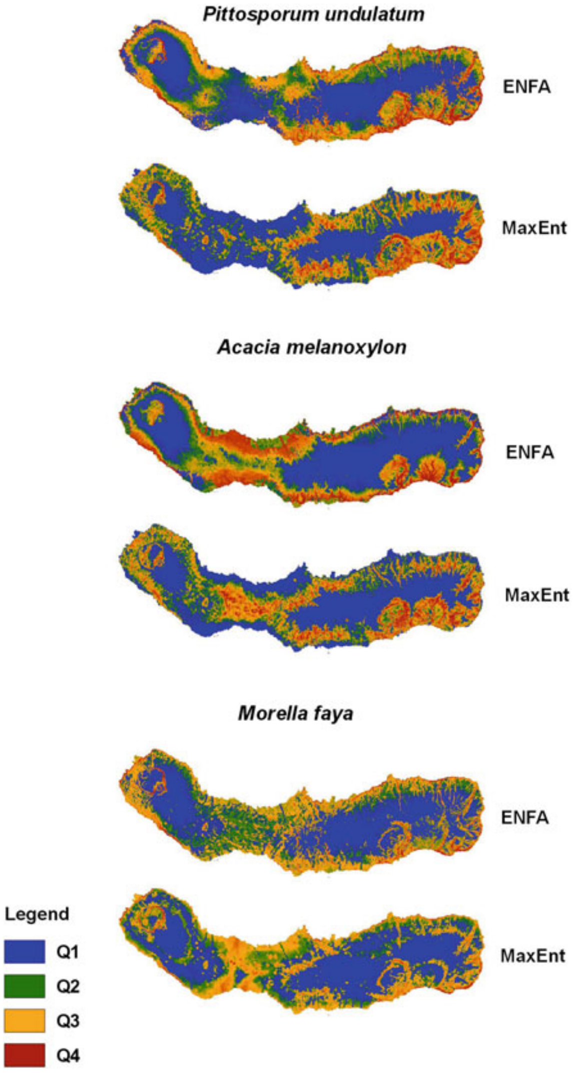


Fig. 10 Habitat suitability maps for *P. undulatum*, *A. melanoxylon* and *M. faya* based on the EGV set 4, derived with occurrence data from São Miguel Island, and applying two modelling approaches, ENFA and MaxEnt. Habitat suitability was reclassified into four levels, according to the respective quartiles

Here we extended those studies by comparing the habitat suitability models derived from ENFA with those derived from MaxEnt. In general, we found that both approaches provided similar results, particularly when the amount of information explained by ENFA was high. Similar results were found by Rupprecht et al. [86].

The modelling results indicated that the digital elevation model, relative humidity and precipitation influence the niche of *P. undulatum*, suggesting that the invader avoids the high mountains, where temperatures are low and relative humidity and rainfall are high.

Table 7 Results of modelling the potential distribution of three woody species in three Azorean islands using four different EGV sets with the ENFA (4 Factors) of MaxEnt. EGV: Ecogeographical variables; ExI: Explained information; BI: Boyce Index; AUC: Area Under the Curve; sd: Standard deviation

Species	Island	EGV set	ENFA			MaxEnt			
			ExI	BI	sd	AUC	sd	BI	sd
P. undulatum	Pico	1	0.948	0.946	0.024	0.689	0.002	0.938	0.041
		2	0.959	0.980	0.008	0.700	0.002	0.977	0.023
		3	0.888	0.982	0.007	0.704	0.002	0.964	0.035
		4	0.878	0.988	0.005	0.708	0.002	0.964	0.045
	Terceira	1	0.905	0.953	0.013	0.799	0.004	0.520	0.161
		2	0.904	0.962	0.027	0.714	0.004	0.898	0.072
		3	0.779	0.928	0.033	0.816	0.006	0.341	0.211
		4	0.784	0.924	0.043	0.849	0.003	0.057	0.203
	São Miguel	1	0.927	0.541	0.088	0.771	0.002	0.271	0.340
		2	0.907	0.996	0.002	0.717	0.001	0.711	0.137
		3	0.779	0.402	0.060	0.778	0.002	0.137	0.437
		4	0.773	0.246	0.147	0.782	0.003	0.198	0.409
A. melanoxylon	Pico	1	0.938	0.947	0.022	0.872	0.006	0.474	0.309
		2	0.954	0.886	0.072	0.874	0.003	0.472	0.279
		3	0.879	0.454	0.165	0.914	0.003	−0.129	0.388
		4	0.869	0.706	0.154	0.916	0.001	−0.083	0.361
	Terceira	1	0.899	0.892	0.026	0.836	0.017	−0.029	0.192
		2	0.931	0.955	0.016	0.791	0.009	0.688	0.131
		3	0.788	0.955	0.025	0.895	0.009	−0.447	0.150
		4	0.768	0.910	0.060	0.906	0.011	−0.507	0.175
	São Miguel	1	0.911	0.885	0.082	0.740	0.002	0.677	0.174
		2	0.903	0.997	0.002	0.697	0.003	0.779	0.101
		3	0.766	0.970	0.029	0.760	0.002	0.577	0.185
		4	0.758	0.992	0.004	0.770	0.002	0.592	0.194
M. faya	Pico	1	0.958	0.946	0.023	0.791	0.002	0.947	0.029
		2	0.971	0.759	0.113	0.793	0.002	0.921	0.046
		3	0.905	0.937	0.034	0.804	0.001	0.885	0.054
		4	0.900	0.938	0.035	0.811	0.001	0.881	0.066
	Terceira	1	0.948	0.844	0.071	0.926	0.001	−0.201	0.221
		2	0.939	0.789	0.102	0.884	0.008	0.285	0.260
		3	0.871	0.829	0.073	0.963	0.002	−0.877	0.111
		4	0.864	0.838	0.071	0.969	0.003	−0.891	0.092
	São Miguel	1	0.958	−0.814	0.045	0.916	0.005	−0.953	0.048
		2	0.887	0.651	0.168	0.846	0.006	0.372	0.162
		3	0.804	−0.804	0.063	0.941	0.004	−0.961	0.037
		4	0.792	−0.806	0.039	0.951	0.004	−0.924	0.072

However, both modelling approaches were somewhat sensitive to a transference of the selected models to a different habitat (i.e., island). Indeed, models of distribution or suitability can be highly sensitive to the definition of the study area [87]. In addition, models are sensitive to issues related to scale. There are no obvious guidelines about which choice of scale is appropriate, because such choice depends on the ecology of the organism at hand and the objectives of the investigation [79, 88]. In our case, we believe that data quality is not a concern, because we used an inventory of the presence of the target species covering the total extent of Pico, Terceira and São Miguel islands.

Furthermore, clear differences between the studied islands seem to occur regarding the present distribution of the target species, particularly for *A. melanoxylon* and *M. faya*, while *P. undulatum* seems to show a more homogeneous distribution at low/intermediate elevations in the three islands. This might be associated with differences in land use, differently affecting the present distribution of the target species. Thus, the situation in each island might differ considerably, not only due to particular environmental conditions, but also to the peculiarities of the species distribution data used for modelling. In fact, besides more or less deterministic events like land use changes associated to gradients of human activity, historic details of the initial colonization as well as chance events might affect the distribution of plant species, originating different starting points for modelling in different regions/islands.

Thus, modelling the distribution in a particular region (i.e. island) might demand a particular model. Alternatively, a global model for the whole region of interest might be obtained (i.e., the archipelago), including all the environmental variation present, but eventually loosing definition of the variables that are more important in each portion of the territory. This should be largely decided based on the objectives of the modelling approach, i.e. a model very adjusted to a particular region, when precise management of a species is required at that area versus a more global model, allowing application to different areas, when only a general prediction of species distribution is required, such as in studies devoted to climate change.

Since using the same EGV set for the different species was not always effective, we should assume that modelling the habitat suitability of the different species making up the exotic woodland in the Azores might require species specific models. This is in agreement with an individualistic view of plant communities, where each species might have particular environmental ranges for the different variables (i.e. a particular ecological niche). This is further complicated since, in the case of invading species, it is possible that a species may behave as a specialist in the early stages of colonization, while becoming more generalist as the population expands [89].

Regarding the more specific results of this research, they are in conformance with the studies cited above [26, 27, 33]. The potential distributions of *P. undulatum* in Pico Island are very similar, using ENFA and MaxEnt algorithms. Although there was a consistent trend in ENFA to perform slightly better. We could confirm a peculiar drawback of MaxEnt when using a large sample. Previous studies confirmed that, in some cases, the performance of MaxEnt increases with small sample sizes, and that it tends to produce restricted predictions [8, 90]. Despite these modelling limitations, the predictions from the different models demonstrated a satisfactory performance in statistical terms. Both the continuous Boyce index values and the shape of the continuous Boyce curves indicated good modelling prediction results when applying a specific sequence of EGV. In particular, EGV set 4 showed the highest value, very close to a maximum of 1. The value reflects how much the model differs from chance expectation or deviation from randomness [34]. However, this measure should be treated with some care, since it is highly dependent on the species niche breadth and on the relevance of the chosen environmental variables.

A number of statistical procedures are available for exploring patterns in presence-only data; the choice among them depends on the quality of the presence only data [91]. We selected this type of data in order to avoid false absences (removal of native species, as *M. faya*) or species that have not reached equilibrium with the environment (invasive species, as *P. undulatum* and *A. melanoxylon*). Presence-only data can provide insight into the vulnerability of species; models developed using these data can inform management [8, 91], however we must be mindful of the biases inherent in the presence data and be cautious in the interpretation of model predictions.

It is important to understand the differences between the two modelling approaches. Tsoar et al. [92] concluded that more complex techniques (e.g., MaxEnt) are better predictors than the simple ones as they establish more flexible relationships between the dependent and independent variables [93]. On the other hand, parametric methods such as ENFA are limited by the normal distribution making them more sensitive to bias [8].

Overall, ENFA assesses the contribution of each variable to the final ecological niche model, allowing an extra validation of the results by an ecologist and conservationist [94]. According to Sérgio et al. [94] this is not possible, at least in a straightforward manner, with MaxEnt. These slight inconsistencies may be due to several factors, primarily data quality and statistical representativeness. It is crucial that the species data provide a non biased sampling of all environmental features where the species occurs. The major difficulty to conciliate the different validation approaches, when applied to MaxEnt outputs, is that there is not an evaluation on which environmental factors weigh more in the model results. ENFA gives this information directly, but has a higher requirement in the number of samples [94]. It is important to highlight that different methods should be used complementary.

Other modelling approaches are also available, including more or less automatic analyses testing different types of modelling algorithms and (biomod2; [95, 96]), modelling approaches based on mathematical concepts [97, 98] or in probabilistic approaches derived within a Bayesian framework [99, 100]. Therefore, future work will be devoted to using the large data set on forest species in the Azores to test and develop different species distribution model alternatives.

References

1. Zaniewski, A.E., Lehmann, A., Overton, J.M.: Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. **157**, 261–280 (2002)
2. Huston, M.A.: Introductory essay: critical issues for improving predictions. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (eds.) Predicting Species Occurrences: Issues of Accuracy and Scale, pp. 7–21. Island Press, Washington (2002)
3. Guisan, A., Zimmermann, N.E.: Predictive habitat distribution models in ecology. Ecol. Model. **135**, 147–186 (2000)
4. Ricklefs, R.E.: A comprehensive framework for global patterns in biodiversity. Ecol. Lett. **7**(1), 1–15 (2004)
5. Graham, C.H., Hijmans, R.J.: A comparison of methods for mapping species ranges and species richness. Global Ecol. Biogeogr. **15**(6), 578–587 (2006)
6. Guisan, A., Thuiller, W.: Predicting species distribution: offering more than simple habitat models. Ecol. Lett. **8**(9), 993–1009 (2005)
7. Franklin, J.: Mapping species distributions. Spatial Inference and Prediction, p. 320. Cambridge University Press, Cambridge (2009)
8. Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.McC., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E.: Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**, 129–151 (2006)
9. Cauwer, V., Muys, B., Revermann, R., Trabucco, A.: Potential realised future distribution and environmental suitability for *Pterocarpus angolensis* DC in southern Africa. For. Ecol. Manag. **315**, 211–226 (2014)
10. Pearce, J., Ferrier, S.: An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecol. Model. **128**, 127–147 (2000)
11. Chytrý, M., Wild, J., Pyšek, P., Jarošík, V., Dendoncker, N., Reginster, I., Pino, J., Maskell, L.C., Vilà, M., Pergl, J., Kühn, I., Spangenberg, J.H., Settele, J.: Projecting trends in plant invasions in Europe under different scenarios of future land-use change. Global Ecol. Biogeogr. **21**(1), 75–87 (2012)
12. Schleupner, C., Link, P.M.: Potential impacts on important bird habitats in Eiderstedt (Schleswig-Holstein) caused by agricultural land use changes. Appl. Geogr. **28**, 237–247 (2008)
13. Falk, W., Mellert, K.H.: Species distribution models as a tool for forest management planning under climate change: risk evaluation of *Abies alba* in Bavaria. J. Veg. Sci. **22**, 621–634 (2011)
14. Henderson, E.B., Ohman, J.L., Gregory, M.J., Roberts, H.M., Zald, H.: Species distribution modeling for plant communities: stacked single or multivariate modeling approaches. Appl. Veg. Sci. **17**, 516–527 (2014)
15. Austin, M.P., Meyers, J.A.: Current approaches to modelling the environmental niche of *Eucalypts*: implications for management of forest biodiversity. For. Ecol. Manag. **85**, 95–106 (1996)
16. Liu, J.: Intergrading ecology with human demography, behavior and socioeconomics: needs and approaches. Ecol. Model. **140**, 1–8 (2001)
17. Forman, R.T.T., Godron, M.: Landscape Ecology, p. 619. Wiley, New York (1986)
18. Pickett, S.T.A., Cadenasso, M.L.: Vegetation dynamics. In: Van der Maarel, E. (ed.) Vegetation Ecology, pp. 172–198. Blackwell Publishing, Malden (2005)
19. Meentemeyer, R.K., Anacker, B., Mark, W., Rizzo, D.M.: Early detection of emerging forest disease using dispersal estimation and ecological niche modeling. Ecol. Appl. **18**, 377–390 (2008)

20. Strubbe, D., Matthysen, E.: Predicting the potential distribution of invasive ring-necked parakeets *Psittacula krameri* in northern Belgium using an ecological niche modelling approach. Biol. Invasions **11**, 497–513 (2009)
21. Wittenberg, R., Cock, M.J.W.: Invasive Alien Species a Toolkit of Best Prevention and Management Practices, p. 228. CAB International, Wallingford (2001)
22. Hulme, P.E.: Trade, transport and trouble: managing invasive species pathways in an era of globalization. J. Appl. Ecol. **46**(1), 10–18 (2009)
23. Mueller-Dombois, D.: Biological diversity and disturbance regimes in island ecosystems. In: Vitousek, P.M., Loope, L.L., Adsersen, H. (eds.) Islands, pp. 163–175. Springer, Berlin (1995)
24. Kueffer, C., Daehler, C.C., Torres-Santana, C.W., Lavergne, C., Meyer, J.Y., Otto, R., Silva, L.: A global comparison of plant invasions on oceanic islands. Perspect. Plant Ecol. Evolut. Syst. **12**(2), 145–161 (2010)
25. Silva, L., Ojeda-Land, E., Rodríguez-Luengo, J.L.: Invasive Terrestrial Flora and Fauna of Macaronesia. Top 100 in Azores, Madeira and Canaries, p. 546. ARENA, Ponta Delgada (2008)
26. Costa, H., Aranda, S., Lourenço, P., Medeiros, V., Azevedo, E.B., Silva, L.: Predicting successful replacement of forest invaders by native species using species distribution models: The case of *Pittosporum undulatum* and *Morella faya* in the Azores. For. Ecol. Manag. **279**, 90–96 (2012)
27. Costa, H., Medeiros, V., Azevedo, E.B., Silva, L.: Evaluating ecological-niche factor analysis as a modeling tool for environmental weed management in island systems. Weed Res. **53**, 221–230 (2013)
28. Caujapé-Castells, J., Tye, A., Crawford, D.J., Santos-Guerra, A., Sakai, A., Beaver, K., Lobin, W., Florens, F.B.V., Moura, M., Jardim, R., Gómes, I., Kueffer, C.: Conservation of oceanic island floras: present and future global challenges. Perspect. Plant Ecol. Evolut. Syst. **12**(2), 107–129 (2010)
29. Silva, L., Elias, R.B., Moura, M., Meimberg, H., Dias, E.: Genetic variability and differentiation among populations of the Azorean endemic gymnosperm *Juniperus brevifolia*: baseline information for a conservation and restoration perspective. Biochem. Genet. **49**(11–12), 715–734 (2011)
30. Silva, L., Smith, C.: A characterization of the non-indigenous flora of the Azores Archipelago. Biol. Invasions **6**(2), 193–204 (2004)
31. Silva, L., Smith, C.: A quantitative approach to the study of non-indigenous plants: an example from the Azores Archipelago. Biodivers. Conserv. **15**(5), 1661–1679 (2006)
32. Lourenço, P., Medeiros, V., Gil, A., Silva, L.: Distribution, habitat and biomass of *Pittosporum undulatum*, the most important woody plant invader in the Azores Archipelago. For. Ecol. Manag. **262**(2), 178–187 (2011)
33. Hortal, J., Borges, P.A.V., Jimenéz-Valverde, A., Azevedo, E.B., Silva, L.: Assessing the areas under risk of invasion within islands through potential distribution modelling: the case of *Pittosporum undulatum* in São Miguel, Azores. J. Nat. Conserv. **18**(4), 247–257 (2010)
34. Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A.: Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. **199**(2), 142–152 (2006)
35. Hirzel, A.H., Le Lay, G.: Habitat suitability modelling and niche theory. J. Appl. Ecol. **45**, 1372–1381 (2008)
36. Mateo, R.G., Felicísimo, A.M., Muñoz, J.: Species distributions models: a synthetic revision. Rev. Chil. Hist. Nat. **84**, 217–240 (2011)
37. Elith, J., Leathwick, J.R.: Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. Ecol. **40**(1), 677 (2009)
38. Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N.: Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology **83**(7), 2027–2036 (2002)
39. Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. Ecol. Model. **190**, 231–259 (2006)
40. Borges, P.A.V., Azevedo, E.B., Borba, A., Dinis, F.O., Gabriel, R., Silva, E.: Ilhas Oceânicas. In: Pereira, H.M., Domingos, T., Vicente, L. (eds.) Portugal Millenium Ecosystem Assessment, pp. 461–508. Lisboa, Celta Editora (2009)

41. França, Z., Cruz, J.V., Nunes, J.C., Forjaz, V.H.: Geologia dos Açores: uma perspectiva actual. Açoreana **10**(1), 11–140 (2005)
42. Nunes, J.C.: A actividade vulcânica na ilha do Pico do Plistocénio Superior ao Holocénio: Mecanismo eruptivo e hazard vulcânico. Tese de doutoramento no ramo de Geologia, especialidade de Vulcanologia, p. 357. Universidade dos Açores, Ponta Delgada (1999)
43. Madeira, J.E., Silveira, A.B.: Active tectonics and first paleoseismological results in Faial, Pico and S. Jorge islands (Azores, Portugal). Ann. Geophys. **46**(5), 761–773 (2003)
44. Madureira, P., Moreira, M., Mata, J., Allègre, C.J.: Primitive néon isotopes in Terceira Island (Azores archipelago). Earth Planet Sci. Lett. **233**, 429–440 (2005)
45. Calvert, A.T., Moore, R.B., McGeehin, J.P., Rodrigues da Silva, A.M.: Volcanic history and ^{40} Ar/^{39} Ar and ^{14}C geochronology of Terceira Island, Azores, Portugal. J. Volcanol. Geotherm. Res. **156**, 103–115 (2006)
46. Moore, R.B.: Volcanic geology and eruption frequency, São Miguel, Azores. Bull. Volcanol. **52**, 602–614 (1990)
47. Dröuet, H.: Catalogue de la Flore des Iles Açores. Bailliere and Fils, Paris (1866)
48. Goodland, T.J., and Healey, R.: The invasion of Jamaican montane rainforests by the Australian tree *Pittosporum undulatum*. School of Agricultural and Forest Sciences, University of Wales, Bangor, p. 54 (1996)
49. Gleadow, R.M., Ashton, D.H.: Invasion by *Pittosporum undulatum* of the forests of Central Victoria. Invasion patterns and plant morphology. Aust. J. Bot. **29**, 705–720 (1981)
50. Dias, E., Araújo, C., Mendes, J.F., Elias, R.B., Mendes, C., Melo, C.: Espécies florestais das ilhas - Açores. In: Silva, J.S. (ed.) Árvores e florestas de Portugal, vol. 6, pp. 199–254. SA/ Fundação Luso-Americana/ Liga para a Protecção da Natureza, Público, Comunicação Social (2007)
51. Moniz, J., Silva, L.: Impact of *Clethra arborea* Aiton (Clethraceae) in a special protection area of São Miguel island, Azores. Arquipél. - Life Mar. Sci. **20A**, 37–46 (2003)
52. Silva, L., Tavares, J.: Phytophagous Insects Associated with Endemic, Macaronesian, and Exotic Plants in the Azores. In: Editorial, Comité (ed.) Avances en Entomologia Ibérica, pp. 179–187. Museo Nacional de Ciencias Naturales (CSIC) y Universidad Autónoma de Madrid, Madrid (1995)
53. Hussain, M.I., González, L., Reigosa, M.J.: Allelopathic potential of *Acacia melanoxylon* on the germination and root growth of native species. Weed Biol. Manag. **11**, 18–28 (2011)
54. Knapic, S., Tavares, F., Pereira, H.: Heartwood and Sapwood Variation in *Acacia melanoxylon* R. Br. trees in Portugal, Forestry **79**, 1–10 (2006)
55. Searle, S.D.: *Acacia melanoxylon*: a review of variation among planted trees. Aust. For. **62**, 79–85 (2000)
56. Silva, L., Tavares, J.: Factors affecting *Myrica faya Aiton* demography in the Azores. Açoreana **8**(3), 359–374 (1997)
57. Schaefer, H.: Chorology and Diversity of the Azorean flora. Dissertationes Botanicae 374, J. Cramer, Stuttgart (2003)
58. Avaliação da Biomassa Disponível em Povoamentos Florestais na Região Autónoma dos Açores (Evaluation of Available Biomass in Forestry Stands in the Azores Autonomic Region). Inventário Florestal da Região Autónoma dos Açores. Direcção Regional dos Recursos Florestais, Secretaria Regional da Agricultura e Florestas da Região Autónoma dos Açores, p. 8 (2007)
59. Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J.: A statistical explanation of MaxEnt for ecologists. Divers. Distrib. **17**, 43–57 (2011)
60. Sheppard, C.S.: How does selection of climate variables affect predictions of species distributions? A case study of three new weeds in New Zealand. Weed Res. **53**, 259–268 (2013)
61. Azevedo, E.B., Pereira, L.S.: Modelling the local climate in island environments: water balance applications. Agric. Water Manag. **40**(2–3), 393–403 (1999)
62. Azevedo, E.B.: Projecto CLIMAAT - Clima e Meteorologia dos Arquipélagos Atlânticos. PIC Interreg IIIB Mac2, 3/A3 (2003)

63. Azevedo, E.B.: Modelação do Clima Insular à Escala Local. Modelo CIELO aplicado à ilha Terceira. Tese de doutoramento no ramo de Ciências Agrárias, Universidade dos Açores, Angra do Heroísmo, p. 247 (1996)
64. Eastman, J.R.: Idrisi: A Grid-Based Geographic Analysis System. Clark University School of Geography, Worcester (1990)
65. Hirzel, A.H., Hausser, J., Perrin, N.: Biomapper 4.0. Lab. of Conservation Biology. Department of Ecology and Evolution, University of Lausanne, Switzerland (2007)
66. Burnham, K.P., Anderson, D.R.: Multimodel inference understanding AIC and BIC in model selection. Sociol. Methods Res. **33**(2), 261–304 (2004)
67. Trabucco, A., Achten, W.M.J., Bowe, C., Aerts, R., Van Orshoven, J., Norgrove, L., Muys, B.: Global mapping of *Jatropha curcas* yield based on response of fitness to present and future climate. Global Change Biol. Bioenergy **2**, 139–151 (2010)
68. MacNally, R.: Regression and model-building in conservation biology, biogeography and ecology: the distinction between - and reconciliation of - predictive and explanatory models. Biodivers. Conserv. **9**, 655–671 (2000)
69. Liang, L., Clark, J.T., Kong, N., Rieske, L.K., Fei, S.: Spatial analysis facilitates invasive species risk assessment. Forest Ecol. Manag. **315**, 22–29 (2014)
70. Song, W., Kim, E., Lee, D., Lee, M., Jeon, S.-W.: The sensitivity of species distribution modeling to scale differences. Ecol. Model. **248**, 113–118 (2013)
71. Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S., Scroggie, M.P., Woodford, L.: Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. J. Appl. Ecol. **48**(1), 25–34 (2011)
72. Hirzel, A.H., Posse, B., Oggier, P.A., Crettenand, Y., Glenz, C., Arlettaz, R.: Ecological requirements of reintroduced species and the implications for release policy: the case of the bearded vulture. J. Appl. Ecol. **41**(6), 1103–1116 (2004)
73. MacArthur, R.: On the relative abundance of bird species. Proc. Natl. Acad. Sci. USA **43**, 293–295 (1957)
74. Phillips, S.J., Dudík, M.: Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography **31**, 161–175 (2008)
75. Phillips, S.J., Dudík, M., Elith, J., Graham, C., Lehmann, A., Leathwick, J., Ferrier, S.: Sample selection bias and presence-only models of species distributions. Ecol. Appl. **19**, 181–197 (2009)
76. Václavík, T., Meentemeyer, R.K.: Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Divers. Distrib. **18**, 73–83 (2012)
77. Peterson, A.T., Papes, M., Soberón, J.: Rethinking receiver operating characteristic analysis applications in ecological niche modelling. Ecol. Model. **213**, 63–72 (2008)
78. Allouche, O., Tsoar, A., Kadmon, R.: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. **43**, 1223–1232 (2006)
79. Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A.: Evaluating resource selection functions. Ecol. Model. **157**(2–3), 281–300 (2002)
80. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, Hewlett Packard Labs (2004)
81. Thompson, M.L.: Assessing the diagnostic accuracy of sequence of tests. Biostatistics **4**, 341–351 (2003)
82. Vaughan, I.P., Ormerod, S.J.: The continuing challenges of testing species distribution models. J. Appl. Ecol. **42**, 720–723 (2005)
83. Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L.: The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography **29**, 773–785 (2006)
84. Baldwin, A.: Use of maximum entropy modeling in wildlife. Res. Entropy **11**, 854–866 (2009)
85. Lobo, J.M., Jiménez-Valverde, A., Real, R.: AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. **17**, 145–151 (2008)
86. Rupprecht, F., Oldeland, J., Finckh, M.: Modelling potential distribution of the threatened tree species *Juniperus oxycedrus*: how to evaluate the predictions of different modelling approaches? J. Veg. Sci. **22**, 647–659 (2011)

87. Soberón, J., Peterson, A.T.: Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers. Inf. **2**, 1–10 (2005)
88. Boyce, M.S., Mao, J.S., Merrill, E.H., Fortin, D., Turner, M.G., Fryxell, J., Turchin, P.: Scale and heterogeneity in habitat selection by elk in Yellowstone National Park. Ecoscience **10**, 321–332 (2003)
89. Jiménez-Valverde, A., Lobo, J.M., Hortal, J.: Not as good as they seem: the importance of concepts in species distribution modelling. Divers. Distrib. **14**, 885–890 (2008)
90. Václavík, T., Meentemeyer, R.K.: Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? Ecol. Model. **220**, 3248–3258 (2009)
91. Pearce, J.L., Boyce, M.S.: Modelling distribution and abundance with presence-only data. J. Appl. Ecol. **43**, 405–412 (2006)
92. Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., Kadmon, R.: A comparative evaluation of presence-only methods for modelling species distribution. Divers. Distrib. **13**, 397–405 (2007)
93. Rebelo, G., Jones, G.: Ground validation of presence-only modelling with rare species: a case study on barbastelles *Barbastella barbastellus* (Chiroptera: Vespertilionidae). J. Appl. Ecol. **47**, 410–420 (2010)
94. Sérgio, C., Figueira, R., Draper, D., Menezes, R., Sousa, A.J.: Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. Biol. Conserv. **135**, 341–351 (2007)
95. Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B.: BIOMOD - a platform for ensemble forecasting of species distributions. Ecography **32**, 369–373 (2009)
96. Vicente, J.R., Fernandes, R.F., Randin, C.F., Broennimann, O., Gonçalves, J., Marcos, B., Pôças, I., Alves, P., Guisan, A., Honrado, J.P.: Will climate change drive alien invasive plants into areas of high protection value? An improved model-based regional assessment to prioritise the management of invasions. J. Environ. Manag. **131**, 185–195 (2013)
97. Rousseeuw, P.J., Struyf, A.: Computing location depth and regression depth in higher dimensions. Stat. Comput. **8**, 193–203 (1998)
98. Cerdeira, J.O., Monteiro-Henriques, T., Martins, M.J., Silva, P.C., Alagador, D., Franco, A.: Mathematical contributions to link biota with environment. J. Veg. Sci. **25**, 1148–1153 (2014)
99. Smolik, M., Dullinger, S., Essl, F., Kleinbauer, I., Leitner, M., Peterseil, J., Stadler, L.M., Vogl, G.: Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. J. Biogeogr. **37**, 411–422 (2009)
100. Kéry, M., Guillera-Arroita, G., Lahoz-Monhort, J.J.: Analysing and mapping species range dynamics using occupancy models. J. Biogeogr. **40**, 1463–1474 (2013)

Using Bayesian Inference to Validate Plant Community Assemblages and Determine Indicator Species

Luís Silva, Flavie Le Jean, José Marcelino and António Onofre Soares

Abstract Recently, we described changes in plant community composition along gradients of anthropogenic disturbance, using a multinomial distribution in a Bayesian framework. Species were organized into categories (e.g. endemic, native, naturalized, invasive) and the proportions of each category in each community were represented by a multinomial vector. We now extend the use of the multinomial distribution to represent all the species in a community, individually, and use this approach to (i) validate plant community assemblages according to their specific composition, and (ii) determine indicator species for each community assemblage. Communities were assembled according to different models: null (all together); saturated (all separated); semi-saturated (only community replicates together); random (random assemblages); gradient (communities assembled in types along an ecological gradient). The models were calculated by using WinBugs and model fit was evaluated using Deviance Information Criterion (DIC). After the best community assemblage was found, we used Bayes rule to estimate the probability of a community, given the presence of a species, and compared the resulting indicator species with those determined by using conventional indicator values (IndVal). Both community assemblage and indicator species analysis gave good results when using two comprehensive plant community

L. Silva (✉) · F. Le Jean
InBIO, Laboratório Associado, CIBIO - Pólo Açores, Departamento de Biologia,
Universidade dos Açores, Ponta Delgada, Açores, Portugal
e-mail: luis.fd.silva@uac.pt

F. Le Jean
e-mail: flavie.le.jean@gmail.com

J. Marcelino · A.O. Soares
CE3C Centre for Ecology, Evolution and Environmental Changes/Azorean
Biodiversity Group, 9501-801 Ponta Delgada, Açores, Portugal
e-mail: jmar06@gmail.com

A.O. Soares
e-mail: antonio.oc.soares@uac.pt

J. Marcelino · A.O. Soares
Departamento de Biologia, Portuguese Platform for Enhancing Ecological Research &
Sustainability (PEERS), Universidade dos Açores, Ponta Delgada, Açores, Portugal

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_21

data sets for the Azores, i.e., a gradient of anthropogenic disturbance and an altitude gradient. Our method allows to (i) statistically validate plant community assemblages; and (ii) incorporate the prevalence of a plant community in the calculations pertaining to indicator species analysis.

Keywords Bayesian analysis · Multinomial distribution · Plant communities · Modelling · Indicator species

1 Introduction

1.1 Biological Communities

Biological communities, i.e. assemblages of individuals belonging to several species that interact in a given area, have been commonly regarded as a supra-specific associations of organisms [1], that are born, develop, grow, and die. However, community structure and dynamics depends not only on species interactions but also on environmental factors, and on historic and stochastic events [2]. Recently, the focus on biological communities has changed from the mere question of delimiting the different communities, to the community-level processes shaping them [3]. In addition, communities may also be regarded as components of a metacommunity [4], with dynamic processes leading to exchange of species among different communities and evolutionary processes occurring within communities.

Determining which processes are most important in shaping communities requires the application of methods that allow to describe and compare them [3]. This is important in order to determine the influence of wide ranging phenomena such as changes in land use, climate change, and biological invasions on biological communities world wide. Grouping of community types has been achieved by using a wide variety of ordination and classification methods [5]. However, other methodologies are required to analyse the various facets of biological communities, namely their functional structure [6]. Here we explore a Bayesian approach to community assemblage, followed by the analysis of indicator species in the community assemblages. We extend a previous analysis where a multinomial distribution was applied to species categories in plant communities, in order to define species spectra and to group community types [7].

1.2 Indicator Species

Identifying indicator species is an often used method in ecology since it improves community survey efficiency due to the capacity of bioindicator species to detect cryptic changes, biotic integrity and sustainability of habitats and farming practices [8–13]. These cost efficient methodologies are increasingly being utilized in detriment of comprehensive community surveys, being commonly reported and widely accepted in monitoring and conservation management schemes [14–18]. Several studies aimed to identify the association strength between one or several species and a community/community assemblage through the use of index values based on species fidelity and specificity [19–21]. In this approach, the identification of one species as indicator in one site is not linked to the presence of other species, hence it is common that too many species are selected as potential indicators and the analyst may have to choose which to use based on the highest indicator value found, and additional criteria such as the frequency of occurrence [19, 22].

1.3 Recent Contributions

Recently, it was possible to describe changes in plant community composition along gradients of anthropogenic disturbance, using a multinomial distribution within a Bayesian framework [7]. The species were organized according to categories (e.g. endemic, native, naturalized, invasive) and the proportions of each category, at each community, were represented by a multinomial vector. This method was also used for the study of vegetation along trails within protected areas [23] and along an altitudinal gradient [24]. Those methods allowed not only to assemble plant communities but also to compare the functional structure of these community assemblages, based on traits such as the life form or the origin/conservation status of the species present at each community.

In addition, we have been using and developing metrics allowing to define indicator species, of particular community types, such as the IndVal method [20] and the ComVal method developed by our team [25]. Here we aim to extend the use of the multinomial distribution to represent all the species in a community, individually, i.e. considering that the categories are the species according to which the individuals are grouped. This approach will be used to (i) group the plant communities in community assemblages, according to their affinity or similarity; and (ii) define indicator species for the different community assemblages.

1.4 Theoretical Background

The multinomial distribution has been used to model different events or processes in ecology, namely population age structure [26] and capture-recapture methods [27], but also in other areas of research [28–32].

The probability function of the multinomial distribution is:

$$f(x_1, \ldots, x_k;\ n;\ p_1, \ldots, p_k) = P(X_1 = x_1\, and \ldots X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} \times p_1^{x_1} \ldots p_k^{x_k} \tag{1}$$

where $\sum_{i=1}^{k} x_i = n$, and x_1 to x_k are non-negative integers.

Let S be the number of species surveyed in C communities. We want to calculate the multinomial vectors representing each community/community assemblage:

$$f_{C_j}(S_1, \ldots, S_k;\ n;\ p_1, \ldots, p_k) = \frac{n!}{S_1! \ldots S_k!} \times p_1^{S_1} \ldots p_k^{S_k} \tag{2}$$

where S_1 to S_k are the numbers of individuals or an estimate of the abundance of each species, and p_1 to p_k are the probabilities of each species in the community/community assemblage, C_j.

That is, each community/community assemblage is represented by a multinomial vector, where each species can be found with a probability, $P(S_i|C_j)$. The probability to find the species, S_i, knowing the community C_j, was computed with a multinomial distribution, using a Dirichlet distribution as prior. This model was also used to assemble the different communities, and we designated this process as BayesCom, since it provides a way to validate community assemblage.

We used the species probabilities derived from each multinomial model, $P(S_i|C_j)$, to calculate an index of similarity. For each community pair (j, h), we calculated the mean overlap in the probabilities for each species, i, as:

$$BayesSim(C_j, C_h) = \frac{\sum \frac{Min[P(S_i|C_j), P(S_i|C_h)]}{Max[P(S_i|C_j), P(S_i|C_h)]}}{S} \tag{3}$$

where S is the number of species. This index has a maximum value of 1, corresponding to the situation where all the species in a pair of communities would have the same probabilities (i.e. the communities would be completely overlapped or similar).

In order to determine indicator species, we then calculate the inverse probability, following the approach generally used to analyse confusion matrices or diagnostic tests, by using Bayes theorem. Accordingly, the probability to identify one community belonging to the group j, knowing that the species i was found is:

$$P(C_j|S_i) = \frac{P(S_i|C_j) \times P(C_j)}{P(S_i)} \tag{4}$$

With $P(S_i)$ the probability to find the species i, being:

$$P(S_i) = \sum P(S_i|C_j) \times P(C_j) \tag{5}$$

So that,

$$P(C_j|S_i) = \frac{P(S_i|C_j) \times P(C_j)}{\sum P(S_i|C_j) \times P(C_j)} \tag{6}$$

Equation (6) provides the indicator value of each species for each community and was designated BayesVal. Here we assume that the communities were sampled at the same number of sites, so $P(C_j)$, the prevalence of the communities, corresponds to one divided by the number of different communities. However, in future applications, the probability of each community type can be used as long as estimates of its prevalence, that is of its frequency or relative extension, are available.

1.5 Objectives

In this research we used this approach, based on the multinomial distribution (Fig. 1), to first validate plant community assemblages (BayesCom, BayesSim) and at a second stage to determine indicator species (BayesVal). Two different datasets concerning Azorean plant communities were used. Finally we compared the results obtained with BayesVal and IndVal. We are more interested in defining methods that allow to detect ecologically meaningful changes in biological communities than in the mere description of those communities as abstract individual entities.

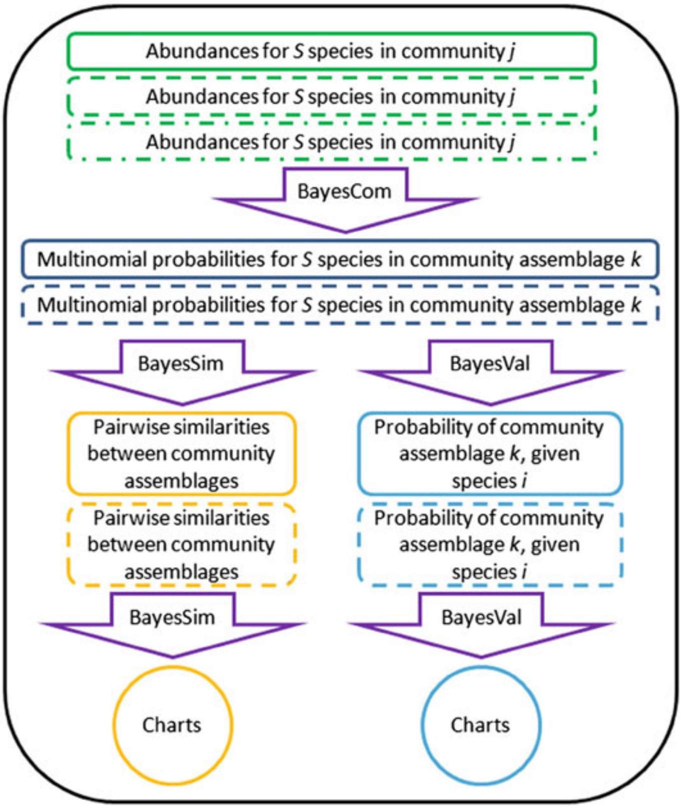

Abundances for S species in community j
Abundances for S species in community j
Abundances for S species in community j
BayesCom
Multinomial probabilities for S species in community assemblage k
Multinomial probabilities for S species in community assemblage k
BayesSim
BayesVal
Pairwise similarities between community assemblages
Pairwise similarities between community assemblages
Probability of community assemblage k, given species i
Probability of community assemblage k, given species i
BayesSim
BayesVal
Charts
Charts

Fig. 1 Flowchart showing a summary of the procedures analysed in this paper to describe plant communities, starting from species abundances and using a multinomial model under a Bayesian framework

2 Methods

2.1 Plant Communities

In this research we used community types along an anthropogenic disturbance gradient [7], and along an altitudinal gradient [24].

Plant Communities Along a Gradient of Anthropogenic Disturbance

We used the dataset presented in [7], where 348 species were recorded along an anthropogenic gradient allowing the identification of nine different arborescent and herbaceous community groups, in five Azorean islands. At each island two replicates per community type were sampled. Due to considerable differences in vegetation structure, we analysed arborescent and herbaceous communities separately. The herbaceous dataset was made of 40 communities, gathering 189 species sampled in five community types: natural meadows, semi-natural pasture at high and low altitude, artificial pasture and corn. The arborescent dataset included 40 communities, gathering 223 species, covering four community types: natural forest, exotic woodland, production forest, and orchards.

Plant Communities Along an Altitudinal Gradient

A data set published in 2014 [24] concerning Azorean plant communities sampled along an altitudinal gradient in a coastal protected area in São Miguel Island, Lombo Gordo. The data set included 36 species among 12 communities corresponding to three replicates at 0–10, 100, 200 and 300 m of altitude.

2.2 Statistical Analysis

Community Assemblages

In order to group the communities into assemblages providing the best fit we run a multinomial model, with each community represented by a multinomial vector with dimension S.

For the anthropogenic disturbance gradient dataset, we have built five models according to different scenarios: (i) a saturated model considering all communities as different; (ii) a semi-saturated model considering that only the replicates of the different communities were considered as similar; (iii) an anthropogenic disturbance gradient model, gathering the replicates of the same type of community, as judged by an expert in Azorean flora and vegetation; (iv) an island model gathering all communities from the same island; and (v) a null model considering that the communities were all similar. We have also built random models by randomly gathering all the communities in each data set into 2, 4, 8, 10 and 20 community assemblages, with 20 replicates in each case.

The same method was applied to the altitudinal gradient dataset, allowing the construction of three models: (i) saturated; (ii) null; and (iii) altitudinal gradient, which gathered the communities sampled at the same altitude.

After checking the validity of the analysis, for each plant community dataset, we considered that the best clustering of plant communities corresponded to the model displaying the best fit.

Model Calculations

We analysed our data using Bayesian inference with the application WinBUGS [33], which has been shown to be an adequate tool for data analysis in ecology [26, 27, 34]. Dirichlet distribution was used as a prior for the multinomial parameters, since it is the conjugate of the probabilities of the multinomial model [26, 27]. This distribution has been successfully used in different Bayesian applications [35, 36]; thus, it has the potential to be used in other modelling approaches, namely in community ecology. In all cases we used three Markov chains and updated the model a number of times that was clearly sufficient to reach chain convergence for the model parameters, by using normally accepted criteria [27], including analysis of trace plots, the Brooks–Gelman–Rubin diagnostic, and the magnitude of Monte Carlo error, as provided by WinBUGS. To estimate model parameters we only considered the estimates obtained after convergence. We used the Deviance Information Criterion (DIC) as a measure of model complexity and fit [37]. In general, we found that updating the model 100,000 times and using the last 30,000 updates to estimate model parameters and DIC, was clearly sufficient to ensure chain convergence. All models were run using the R package "R2WinBUGS" [38].

BayesSim

An R script was prepared to calculate BayesSim from the probability values, $P(S_i|C_j)$, derived from WinBUGS. The similarities between the communities were represented using chord diagrams edited using the R package "circlize" [39].

BayesVal and IndVal

The model with the best fit was then used to compute the probability of one site to belong to a target community group, knowing the presence of one species, $P(C_j|S_i)$. Those species having their BayesVal above 0.5 were selected as indicator species for further analyses, since they would have a larger indicator value than a random binomial process with a success rate of 0.5. At the same time, we calculated the conventional species indicator value (IndVal) based on the fidelity and specificity of species in relation to plant communities, for each species/community assemblage, applying the "multipatt" function of the "indicspecies" R package [40], and we compared the results obtained with both approaches.

3 Results

3.1 Assemblage of Plant Communities

For both data sets, the null models obtained the highest DIC which proves that there was a detectable variation in species composition among the communities (Tables 1, 2). The lowest DIC value was obtained with the gradient models for both arborescent

Table 1 Model fit obtained for different multinomial models adjusted to two data sets (Arborescent and Herbaceous) regarding plant communities along a gradient of anthropogenic disturbance. DIC values were obtained after convergence, by using three Markov Chains, 100,000 updates, the first 70,000 discarded

Model	Arborescent	Herbaceous
Null	8851.9	8111.6
Island	8132.3	7141.9
Saturated	7508.0	8111.4
Semi-saturated	7151.6	6271.4
Gradient	6936.1	5973.7
Random (mean of 20 replicates)		
20 community groups	8602.7	7863.0
10 community groups	8241.0	7577.9
8 community groups	7956.2	7297.7
4 community groups	7871.6	7224.6
2 community groups	7639.5	7059.1

Table 2 Model fit obtained for different multinomial models adjusted to a data set regarding plant communities along an altitudinal gradient (0–10, 100, 200 and 300 m). DIC values were obtained after convergence, by using three Markov Chains, 100,000 updates, the first 70,000 discarded. DIC Value of models for plant community datasets along an altitudinal gradient (0–10, 100, 200 and 300 m). Values obtained after model convergence (three Markov chains with 100,000 updates, the first 70,000 being discarded) by running different multinomial models

Model	Altitudinal
Null	658.3
Saturated	542.5
Gradient	489.0
Random (mean of 20 replicates)	
6 community groups	614.4
4 community groups	596.4
3 community groups	575.0
2 community groups	559.5

and herbaceous communities and for the altitudinal gradient dataset. Regarding random models, the larger the number of community groups randomly created, the larger was the DIC value (Tables 1, 2). Since the DIC values for the gradient models were consistently lower than for all the other models, including the large set of random models, based on different numbers of community groups, we considered that the gradient models best represented the community assemblages in the two data sets. We then estimated the indicator species for each type of community assemblage, in four community groups for the arborescent gradient, in five community groups for the herbaceous gradient, and in four community groups for the altitudinal gradient.

In Fig. 2 it is possible to visualize the similarities between the validated community assemblages. Regarding the altitudinal gradient (Fig. 2, top chart), it seems clear that the coastal communities (0–10 m) showed the lowest similarities towards the other sampled communities, while two types of exotic woodland found at 200 and 300 m showed the highest similarity values. Concerning the arborescent gradient (Fig. 2, middle chart), *Cryptomeria japonica* woodland and invasive woodland showed the highest similarity level. Finally, regarding the herbaceous gradient (Fig. 2, bottom chart), similarities between the communities were generally lower than in the previous cases, with the highest similarity between natural meadows and semi-natural pasture at high elevation.

3.2 Determination of Indicator Species

It was possible to determine indicator species for each of the defined community types, for each of the three gradients. We only consider here those species that showed a p value below 0.05 for IndVal, and those that showed a BayesVal above 0.5. The former p value is provided by the “indicspecies” R package, and results from a permutation test.

Regarding the altitudinal gradient, the number of indicator species was somewhat larger for BayesVal than for IndVal (Fig. 3). However, as discussed below, the indicator species determined by BayesVal are ecologically meaningful. For instance, the invader *Acacia melanoxylon* is commonly found at exotic woodlands, sometimes dominating the biological communities. Here, it appears as an indicator of one of the community assemblages, an exotic woodland. The introduced *Aloe arborescens* is commonly found at coastal areas where it tends to spread vegetatively. *Picconia azorica*, an endemic tree, defines a peculiar type of natural coastal woodland. Thus, those species, and several others in Fig. 1, show the transition form exotic woodland to a native woodland and to coastal vegetation, as the altitude decreases. Other examples include *Asplenium marinum*, *Plantago coronopus* and *Spergularia azorica*, clearly associated with the coastal areas in the Azores.

Considering the herbaceous and the arborescent anthropogenic gradients (Figs. 4 and 5), in general, the number of indicator species was similar for both BayesVal and IndVal metrics, with a considerable overlap of the indicators species defined by the two methods for some of the communities, namely pasture, semi-natural pasture at low elevation, natural meadows, orchards and natural forests. In general, the indicator species were also ecologically meaningful, as several fruit tree and hedgerow species in orchards (e.g. *Anona cherimolia*, *Banksia integrifolia*, *Citrus* spp.), or several indigenous species in natural forests, including trees/shrubs (e.g. *Erica azorica*, *Ilex perado* ssp. *azorica*, *Myrsine africana*) and ferns (*Dryopteris* spp.). For instance in corn, three other crops often found in association, namely *Cucurbita pepo*, *Ipomoea batatas* and *Phaseolus vulgaris*, were determined as indicator species by BayesVal. In natural meadows, several endemic grasses were found to be indicator species, namely *Deschampsia foliosa* and *Festuca francoi*. In semi–natural pasture at low

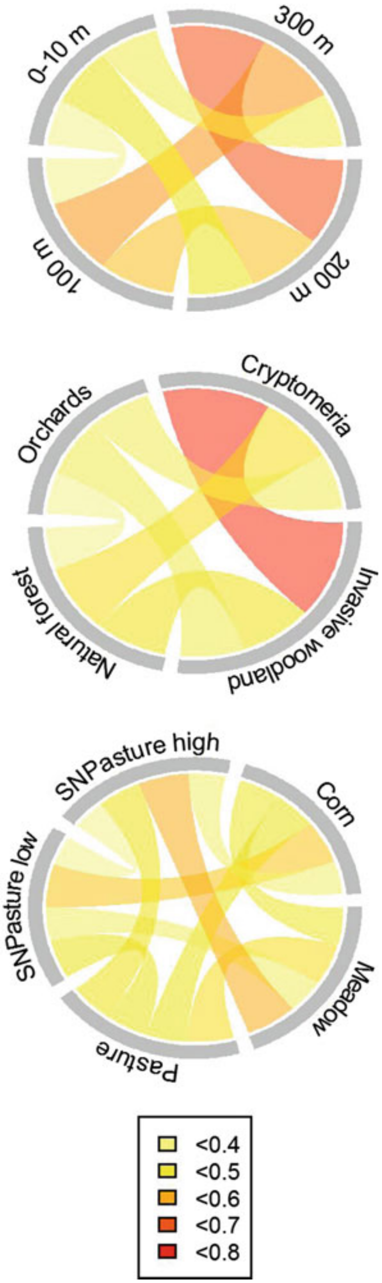


Fig. 2 Similarities between community assemblages for two data sets, one regarding an altitudinal gradient (*top chart*) and the other an anthropogenic gradient (*middle chart*, herbaceous communities; *bottom chart*, arborescent communities). The similarity index, BayesSim, was based on species probabilities, calculated by using a multinomial model

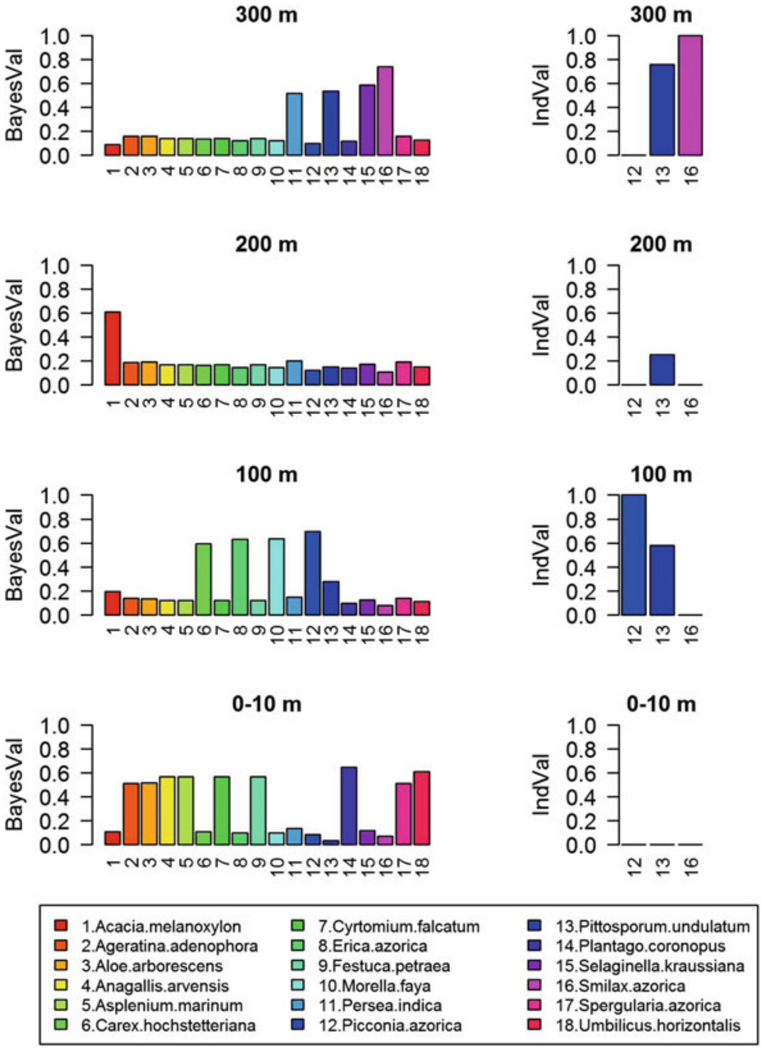


Fig. 3 Comparison of the indicator species determined by using BayesVal or Indval, along an altitudinal gradient. Only those species with BayesVal above 0.5 and with the *p* value for IndVal below 0.05 are shown

elevation *Daucus carota* is a common component while, at high elevation, several indigenous taxa emerge as indicators, namely *Hidrocotyle vulgaris* and *Selaginella kraussiana*, typical of humid locations, and *Veronica officinalis*, common in this type of pastures.

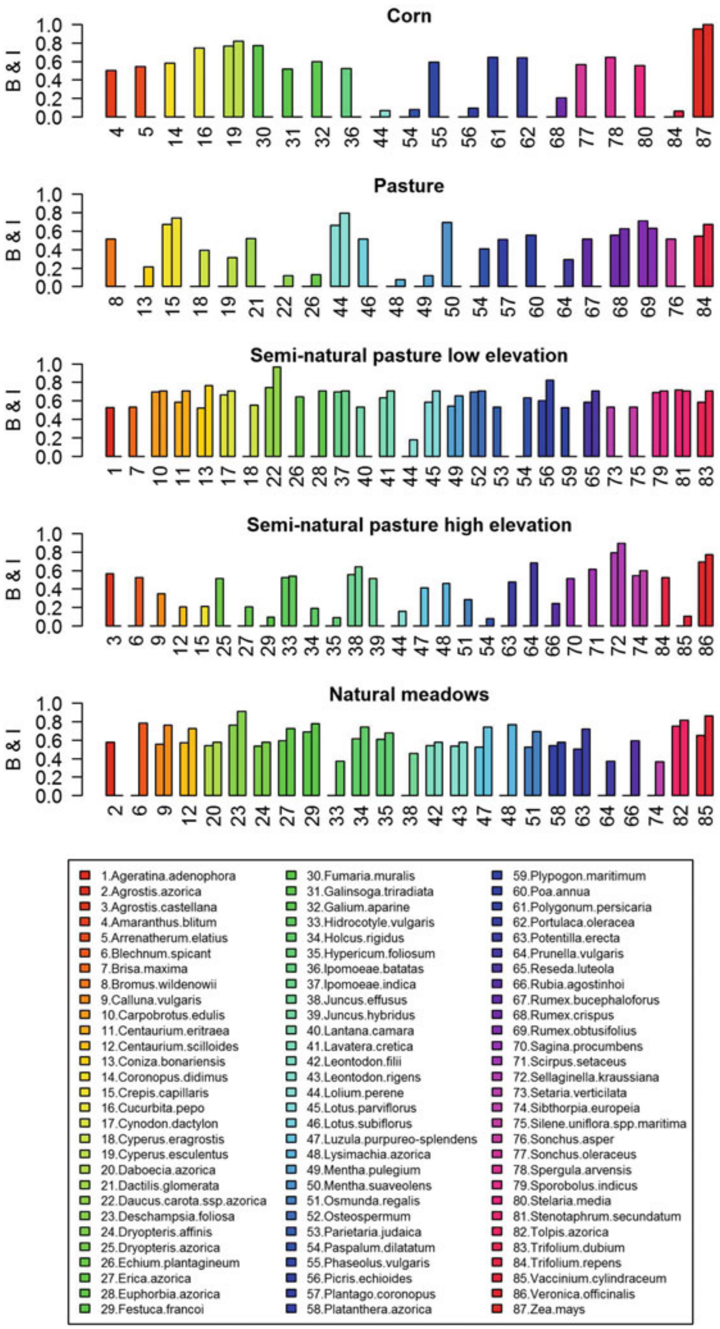


Fig. 4 Comparison of the indicator species determined by using BayesVal or Indval, for herbaceous communities along an anthropogenic gradient. Only those species with BayesVal above 0.5 and with the p value for IndVal below 0.05 are shown. In the barplot, for each species, the first (*left*) column reports BayesVal while the second (*right*) column reports IndVal

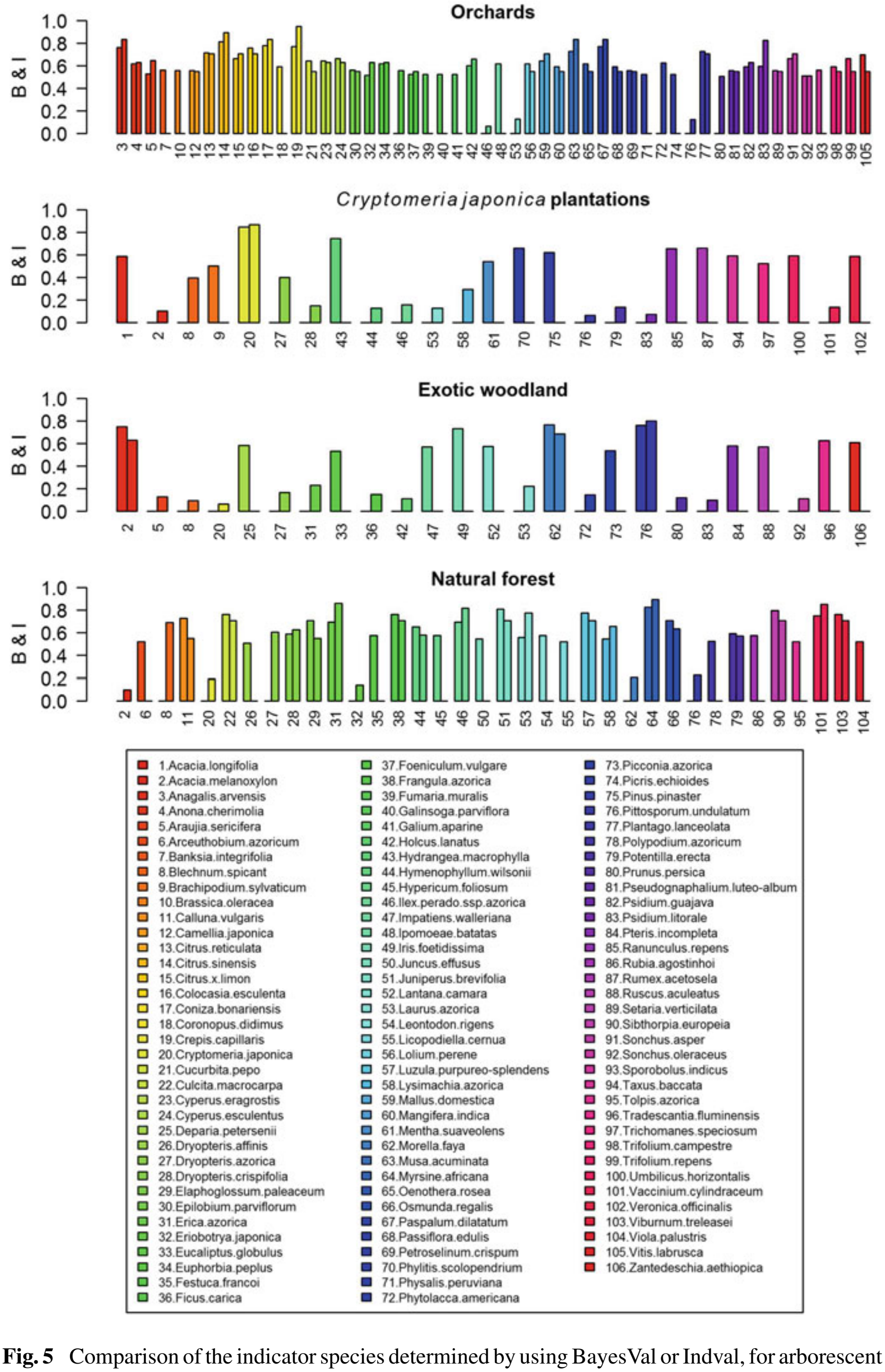


Fig. 5 Comparison of the indicator species determined by using BayesVal or Indval, for arborescent communities along an anthropogenic gradient. Only those species with BayesVal above 0.5 and with the p value for IndVal below 0.05 are shown. In the barplot, for each species, the first (*left*) column reports BayesVal while the second (*right*) column reports IndVal

4 Discussion

Results regarding the altitudinal gradient were able to clarify previous analyses using a Bayesian multinominal approach to plant community assemblage. When we first analysed this data set [24], it was not possible to clearly discriminate between the communities along the altitudinal gradient, when the approach was based on functional groups (i.e. species divided according to their origin/status or to their life–form). However, using the new approach listed herein, and data at the species level, the altitudinal model was clearly the best, showing the best fit. Thus, moving the analysis from the level of species categories to the species level increased the discriminating power of the method. Considering the indicator species, BayesVal showed relevant differences when compared to IndVal. When analysing this data set, the number of indicator species defined by using IndVal was lower than when using BayesVal. Nonetheless, the interpretation of the BayesVal indicator species is straightforward and concurs with current literature on the distribution of species in the Azores, making this metric ecologically meaningful (Fig. 1) [41–45]: (i) at 300 m we have an exotic woodland with *Persea indica* and *Pittosporum undulatum* standing out as well as the endemic vine, *Smilax azorica*; (ii) at 200 m, in the exotic woodland *Acacia melanoxylon* presents de highest BayesVal; (iii) at 100 m we have a natural woodland with several indigenous species as indicators; and at 0–10 m we have several indigenous (e.g. *Spergularia azorica*) and non-indigenous species (e.g. *Cyrtomium falcatum*) that commonly occur in the coastal areas in the Azores.

Regarding the herbaceous and the arborescent communities, the validation of the community types according to the position along the gradient was the same as that already found when analysing species grouped in categories [7]. Concerning the indicator species, when considering the arborescent gradient, BayesVal and IndVal largely coincided for the extreme communities (i.e. orchards and natural forest), with a smaller overlap for the intermediate communities. In the latter case, there was overlap for those species characterizing the communities, namely *Cryptomeria japonica*, *Acacia malanoxylon*, *Pittosporum undulatum* and *Morella faya*. When considering the herbaceous gradient, there was a considerable overlap between BayesVal and IndVal regarding natural meadows and semi-natural pastures at low elevation. Concerning pasture, besides species also indicated by IndVal, BayesVal also selected other species characteristic of the Azores pastures, including *Bromus wildenowii* and *Mentha suaveolens* [46]. At the semi-natural pastures at high elevation, BayesVal indicated a very important species at those types of plant communities in the Azores, *Agrostis castellana* [46]. Interestingly, several of the species indicated by IndVal but not by BayesVal, were considered as indicators in more than one community, namely *Blechnum spicant*, *Cyperus eragrostis* and *Prunella vulgaris*. Still in other cases (e.g. *Coniza bonariensis*, *Crepis capillaris*, *Lolium perene*, *Vaccinium cylindraceum*), BayesVal only coincided with IndVal for the community where the latter showed the largest value (semi-natural pasture low elevation, pasture, pasture, and natural meadows, respectively).

In some cases, the statistical process underlying the calculation of BayesVal seems to make it susceptible to select indicator species that, although of seldom occurrence, appeared in the sampled metacommunity associated with one community type only, as was the case of *Hydrangea macrophylla* towards *Cryptomeria japonica* woodland. As in other methodologies, it still depends on the researcher to discern when the occurrence of a particular species is only a sporadic event, resulting not from a meaningful association but from a casual phenomena: hydrangeas are common in the margins of *Cryptomeria japonica* stands, thus it is not unlikely that some of the shrubs manage to grow inside the forest stand.

Globally, however, both BayesCom and BayesVal allowed to assemble plant communities and to select indicator species in a ecologically meaningful way. Furthermore, if the prevalence of the several types of communities is available, BayesVal can take that information into account. Thus, this statistical method might be regarded as a unifying approach, based on a multinomial model that can be used to validate community assemblages, represent the affinities between community groups, and determine indicator species, therefore providing a comprehensive characterization of biological communities within a given metacommunity.

References

1. Clements, F.E.: Nature and structure of the climax. J. Ecol. **24**, 252–284 (1937)
2. Gleason, H.A.: The individualistic concept of the plant association. Bull. Torrey Botanical Club **53**, 7–26 (1926)
3. Gurevitch, J., Scheiner, S.M., Fox, G.A.: The Ecology of Plants. Sinauer Associates Inc Publishers, Sunderland (2002)
4. Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M., Gonzalez, A.: The metacommunity concept: a framework for multi-scale community ecology. Ecol. Lett. **7**, 601–613 (2004)
5. Borcard, D., Gillet, F., Legendre, P.: Numerical Ecology with R. Springer, New York (2011). 306 pp
6. Mouillot, D., Villger, S., Scherer-Lorenzen, M., Mason, N.W.H.: Functional structure of biological communities predicts ecosystem multifunctionality. PLoS ONE **6**(3), e17476 (2011). doi:10.1371/journal.pone.0017476
7. Marcelino, J.A.P., Silva, L., Garcia, P.V., Weber, R., Soares, A.O.: Using species spectra to evaluate plant community conservation value along a gradient of anthropogenic disturbance. Environ. Monit. Assess. **185**, 6221–6233 (2013)
8. Paoletti, M.G.: Using bioindicators based on biodiversity to assess landscape sustainability. Agric. Ecosyst. Environ. **74**, 1–18 (1999)
9. Niemi, G.J., McDonald, M.E.: Application of ecological indicators. Annu. Rev. Ecol. Evol. Syst. **35**, 89–111 (2004)
10. Hodkinson, I.D., Jackson, J.K.: Terrestrial and aquatic invertebrates as bioindicators for environmental monitoring, with particular reference to mountain ecosystems. Environ. Manag. **35**, 649–666 (2005)
11. McGeoch, M.A.: Bioindicators. Encyclopedia of Environmetrics. Wiley, London (2006)
12. Muramoto, J., Gliessman, S.R.: Use of bioindicators for assessing sustainability of farming practices. In: Pimentel, D. (ed.) Encyclopedia of Pest Management, pp. 37–41. Marcel Dekker, Taylor & Francis Group, Boca Raton, FL (2006)

13. Odland, A.: Interpretation of altitudinal gradients in South Central Norway based on vascular plants as environmental indicators. Ecol. Indic. **9**, 409–421 (2009)
14. Nordén, B., Paltto, H., Gtmark, F., Wallin, K.: Indicators of biodiversity, what do they indicate? - Lessons for conservation of cryptogams in oak-rich forest. Biol. Conserv. **135**, 369–379 (2007)
15. Billeter, R., Liira, J., Bailey, D., et al.: Indicators for biodiversity in agricultural landscapes: a pan European study. J. Appl. Ecol. **45**, 141–50 (2008)
16. Marignani, M., Vico, E., Maccherini, S.: Performance of indicators and the effect of grain size in the discrimination of plant communities for restoration purposes. Community Ecol. **9**, 201–206 (2008)
17. Winter, S.: Forest naturalness assessment as a component of biodiversity monitoring and conservation management. Forestry **85**, 293–304 (2012)
18. Vilches, B., De Cáceres, M., Snchez-Mata, D., Gaviln, R.G.: Indicator species of broad-leaved oak forests in the eastern Iberian Peninsula. Ecol. Indic. **26**, 44–48 (2013)
19. De Cáceres, M., Legendre, P.: Associations between species and groups of sites: indices and statistical inference. Ecology **90**, 3566–3574 (2009)
20. De Cáceres, M., Legendre, P., Wiser, S.K., Brotons, L.: Using species combinations in indicator value analyses. Methods Ecol. Evol. **3**, 973–982 (2012)
21. Urban, N.A., Swihart, R.K., Malloy, M.C., Dunning, J.B.: Improving selection of indicator species when detection is imperfect. Ecol. Indic. **15**, 188–197 (2012)
22. Dufrene, M., Legendre, P.: Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. **67**, 345–366 (1997)
23. Queiroz, R.E., Ventura, M.A., Silva, L.: Plant diversity in hiking trails crossing Natura 2000 areas in the Azores: implications for tourism and nature conservation. Biodivers. Conserv. **23**, 1347–1365 (2014)
24. Prestes, A., Magalhes, B., Xavier, E., Silva, L.: Changes in plant community composition along an altitudinal gradient on a coastal protected area in the Azores: a Bayesian analysis. Silva Lusit. **special number**, 91–114 (2014)
25. Marcelino, J.A.P., Weber, R., Silva, L., Garcia, P.V., Soares, A.O.: Expedient metrics to describe plant community change across gradients of anthropogenic influence. Environ. Manag. **54**, 1121 (2014). doi:10.1007/s00267-014-0321-z
26. McCarthy, M.A.: Bayesian Methods for Ecology. Cambridge University Press, Cambridge (2007)
27. King, R., Morgan, B.J.T., Gimenez, O., Brooks, S.P.: Bayesian Analysis for Population Ecology. Chapman & Hall, Boca Raton (2010)
28. Boender, C.G.E.: Rinnooy, Kan, A. H. G.: A multinomial Bayesian approach to the estimation of population and vocabulary size. Biometrika **74**, 849–856 (1987)
29. Vasko, K., Toonen, H.T.T., Korhola, A.: A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. J. Paleolimnol. **24**, 243–250 (2000)
30. Griffiths, T.L., Tenenbaum, J.B.: Using vocabulary knowledge in Bayesian multinomial estimation. Adv. Neural Inf. Process. Syst. **14**, 1385–1392 (2002)
31. Kazembe, L., Namangale, J.: A Bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. Eur. J. Epidemiol. **22**, 545–556 (2007)
32. Calvo, E.: The competitive road to proportional representation. World Polit. **61**, 254–295 (2009)
33. Spiegehalter, D.J., Thomas, A., Best, N.G.: WinBUGS User Manual, Version 14. MCR Biostatistics Unit, Cambridge (2003)
34. Kéry, M.: Introduction to WinBUGS for Ecologists. A Bayesian Approach to Regression, ANOVA, Mixed Models and Related Analyses. Academic Press, Burlington (2010)
35. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155**, 945–959 (2000)
36. Huelsenbeck, J.P., Ronquist, F.: MRBAYES: Bayesian inference of phylogeny. Bioinformatics **17**, 754–755 (2001)
37. Spiegehalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. J. R. Stat. Soc.: Ser. B **64**, 583–639 (2002)

38. Sturtz, S., Ligges, U., Gelman, A.: R2WinBUGS: a package for running WinBUGS from R. J. Stat. Softw. **12**, 1–16 (2005)
39. Gu, Z., Gu, L., Eils, R., Schlesner, M., Brors, B.: Circlize implements and enhances circular visualization in R. Bioinformatics **30**, 2811–2812 (2014)
40. De Cáceres, M., Jansen, F.: Package indicspecies. Studying the statistical relationship between species and groups of sites (2014) CRAN. http://cran.r-project.org/web/packages/indicspecies/indicspecies.pdf. 15 Oct 2014
41. Silva, L.: Exotic Woodland In: Cardoso P., Gaspar C., Borges, P.A.V., Gabriel, R., Amorim, I.R., Martins, A.F., Maduro-Dias, F., Porteiro,J.M., Silva, L., Pereira, F. (eds.) Azores – a Natural Portrait/Açores – Um Retrato Natural, pp. 146–151. Veraçor, Ponta Delgada (2009)
42. Lourenço, P., Medeiros, V., Gil, A., Silva, L.: Distribution, habitat and biomass of Pittosporum undulatum, the most important woody plant invader in the Azores Archipelago. For. Ecol. Manag. **262**, 178–187 (2011)
43. Sjögren, E.: Recent changes in the vascular flora and vegetation of the Azores islands. Memórias da Sociedade Broteriana **22**, 1–453 (1973)
44. Dias, E., Elias, R.B., Melo, C., Mendes, C.: Biologia e ecologia das florestas das ilhas. In: Silva, J.S. (ed.) Açores – Árvores e florestas de Portugal, Açores e Madeira. A floresta das ilhas, pp. 51–80, Público Comunicação Social, SA e Fundação Luso–Americana para o Desenvolvimento (2007)
45. Schäfer, H.: Chorology and diversity of the Azorean flora. Dissertationes Botanicae **374**, 1–130 (2003)
46. Oliveira, J.N.B.: A pastagem permanente da ilha de São Miguel (Açores): Estudo fitossociológico, fitoecológico e primeira abordagem do ponto de vista agronómico. Dissertação de Doutoramento, Universidade dos Açores, Ponta Delgada, 366 pp. (1989)

Development of Allometric Equations for Estimating Above-Ground Biomass of Woody Plant Invaders: The Case of *Pittosporum undulatum* in the Azores Archipelago

Lurdes Borges Silva, Patrícia Lourenço, Nuno Bicudo Ponte, Vasco Medeiros, Rui Bento Elias, Mário Alves and Luís Silva

Abstract The use of biomass for energy production has shown a growing interest in recent years. *Pittosporum undulatum* is a widespread invasive tree in the Azores archipelago with a considerable potential as an energy resource. Sustainable use of forest resources demands accurate and precise estimation of standing biomass but for

L. Borges Silva (✉) · L. Silva
InBIO, Rede de Investigação em Biodiversidade, Laboratório Associado,
CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Polo-Açores,
Departamento de Biologia, Universidade dos Açores, 9501-801 Ponta Delgada, Açores,
Portugal
e-mail: lurdes.cb.silva@uac.pt

L. Silva
e-mail: luis.fd.silva@uac.pt

P. Lourenço
Andalusian Center for the Assessment and Monitoring of Global Change (CAESCG),
Deparment of Biology and Geology, University of Almera, Almera, Spain
e-mail: pmrlourenco@gmail.com

N.B. Ponte
Azorina SA-Sociedade de Gestão Ambiental e Conservação da Natureza, Avenida Antero de
Quental, No9C, 2oandar, Edifício CTT, 9500-160 Ponta Delgada, Portugal
e-mail: Bikudo10@gmail.com

V. Medeiros
Direção Regional dos Recursos Florestais dos Açores, Rua do Contador, 23, 9500-050 Ponta
Delgada, Açores, Portugal
e-mail: Vasco.AM.Medeiros@azores.gov.pt

R.B. Elias
CE3C, Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group,
Departamento de Ciências Agrárias, Universidade dos Açores, Angra do Heroísmo, 9700-042
Açores, Portugal
e-mail: rui.mp.elias@uac.pt

M. Alves
CEO NATURALREASON, Lda, Caminho do Meio Velho, Cabo da Praia, 5-B, 9760-114
Açores, Portugal
e-mail: mario.alves@naturalreason.pt

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_22

the Azores only a few studies exist devoted to estimation of above-ground biomass (*AGB*). Here we used published allometric equations and developed new models for the estimation of *P. undulatum* total *AGB* in the exotic woodlands of the Azores. A total of 470 trees were sampled at 11 *P. undulatum* stands in São Miguel Island, and their total biomass and several dendrometric traits were recorded in the field. Several dendrometric variables were used as predictors for biomass: diameter at breast height, *D*; tree height,*H*; basal area, *BA*; canopy height, *CH*; canopy diameter, *CD*; canopy biovolume, *BV*; and number of branches at breast height, *NB*. For model comparison and evaluation R^2, *RMSE* and *AIC* were used. The equations comprised models fitted to log-transformed data. There were significant and strong linear relationships between *AGB* and predictors *BA* and *H*. The considerable variability in *P. undulatum* stands, in terms of tree density and structure, should be considered when estimating biomass with allometric equations.

Keywords Pittosporum undulatum · Above ground biomass · Linear regression · Allometric equation

1 Introduction

An innovated interest in tree biomass studies has been documented by the scientific community [1–3]. Standing above-ground biomass (*AGB*) is a fundamental state variable (i.e. a variable used to describe the mathematical state of a dynamical system [4]) in the evaluation of several ecological and eco-physiological models [5], namely for the sustainable management of forest resources, to evaluate the effects of climate change, and to quantify the biomass obtained from forests as an alternative to fossil fuels [3]. Tree biomass estimation is needed for the sustainable planning and management of forest resources. Thus the development of biomass equations has been, and indeed remains, one of the main lines of work of many researchers [2]. The determination of forest biomass can be achieved through several methods. The two commonly used include [6]: (i) the destructive harvesting of plants followed by the determination of their biomass, and (ii) the use of allometric equations to predict plant biomass. Weighing trees in the field is undoubtedly the most accurate method of estimating aboveground tree biomass but it is time-consuming. Generally this method is based on small sample sizes [1], what generates considerable uncertainty when the derived estimates are extrapolated to larger areas [7]. So, it might be preferable to use allometric equations as they allow the estimation of forest biomass for large areas while avoiding forest destruction [8].

Allometric equations have been developed for different purposes, regions and species because tree species may differ greatly in tree architecture [1]. Several modeling approaches have been used, including species-specific allometric models [9], generalized allometric equations [10], simplified allometric models [11], allometric models for regional and global level biomass estimations [12], as well as studies on the uncertainty of using allometric models [13]. This type of analysis implies the use of appropriate statistical techniques to correct the heterogeneity of the variance of the residuals and also an adjustment to ensure additivity between the estimated biomass equations of the different components and the total tree biomass equation [2]. These models were developed using regression techniques for specific geographic areas and tree species, being effective for specific purposes. There is no single optimal model which can provide a good calibration function for the estimation of *AGB* for all tree species and climatic regions because the calibration coefficients of allometric models are reported to vary with tree species, stand age, site quality and climate [14] stage of the forest, disturbance levels, species composition, etc. [15, 16]. Despite all that variation, the equations are generated from a small sample of trees and are then used to estimate biomass on a larger scale.

Allometric equations relate tree biomass (kg) or stand biomass (Mgha^{-1}), as well as their components, with easily measurable variables. Tree measures commonly used, as variables for model building, are diameter at breast height (*D*), total tree height (*H*), and basal area (*BA*), the latter is frequently used as a surrogate for biomass and carbon in tropical moist and dry forests [17]. The *BA* is a good predictor for biomass and carbon since it integrates the effect of both the number and size of trees, i.e. the sum of cross-sectional area measured at breast height (1.3 m) of all trees in a stand, expressed as m^{2}ha^{-1} [18]. Thus, a correlation between *BA* and *AGB* it is to be expected since these variables are both related to the trunk diameter [19].

To contribute to the sustainable management of the exotic woodland in the Azores, the present study focused on *Pittosporum undulatum* Ventenat (Pittosporaceae), a tree or shrub native of Australia introduced in the Azores Islands in the 19th century [20]. This species was ranked eighth, in a total of 195 evaluated species, by an assessment of the Top 100 invasive species in Macaronesia [21]. *P. undulatum* has been able to colonize a wide range of habitats in the nine Azorean islands in less than a century [22–24]. According to a random survey of vascular plants in the Azores, the invader is present throughout the archipelago, occupying 49% of the forested area, approximately 24,000 ha. Further, in previous studies in the Azores, the annual biomass production of *P. undulatum* was quantified, ranging from only about 150 Mg in the small Island of Corvo up to more than 60,000 Mg in Pico Island [22]. However, new species-specific allometric equations are necessary to achieve higher levels of accuracy when estimating *P. undulatum* standing biomass in the Azores.

The goal of this study was to estimate *AGB* of *P. undulatum* trees in exotic woodland in São Miguel Island by using tree allometric equations. Our purpose was to develop new equations by focusing on above-ground woody biomass rather than on total above-ground biomass, because woody organs (stump, stem and branches) contain the majority of the biomass. Our specific objectives were:

- To establish the relationship between *AGB* and tree parameters such as diameter at breast height *D*, tree height *H*, basal area *BA*, canopy height *CH*, canopy diameter *CD*, canopy biovolume *BV*, and number of branches at breast height *NB*;
- To compare our derived models with those frequently used;
- To explore how the selection of allometric equations influences regression modeling (e.g., variable selection, variation explained).

2 Materials and Methods

2.1 Study Area

The study area is located in São Miguel Island, Azores, the largest island of the archipelago, with a surface area of 745 km^2, and the highest peaks at 1105 m above sea level. The Azores archipelago is located in the North Atlantic Ocean between 36° 55'N and 39°42'N and 25°00'W and 31°30'W, about 1500 km west of mainland Portugal. With a total surface area of 2323 km^2, it includes nine volcanic islands, spanning 615 km from east to west [25].

The Azores climate is influenced by the position of the archipelago, which lies in an open oceanic basin, open to the circulations from the North Pole and the tropics [26]. Therefore, in the Azores islands marine air masses interact with cold or temperate air masses from the Pole. Nevertheless, in addition to this constraint, imposed by the Azores position in a global circulation context, local factors, such as distance from coast, altitude and the exposition of the island slopes, also influence the climate [27]. The Azores climate can be considered as marine temperate, which is reflected by low thermal amplitude, high precipitation, high air humidity and persistent wind. One of main characteristics of the Azores climate is the well-established difference between a dry season and a colder, wet season [26].

The Azores islands are characterized by a low number of endemic species compared with the neighboring archipelagos of Madeira and Canaries [28]. This fact may be due to the long distance to the nearest mainland (Europe, 1300 km), low geological age and the homogeneous oceanic climate of the islands. The natural plant communities, existing prior to human settlement, include coastal and wetland vegetation, meadows, and various types of scrubland and forest [29]. However, the native vegetation on São Miguel Island is threatened by several aggressive invaders, including *P. undulatum*, *Cryptomeria japonica*, *Hedychium gardnerianum*, *Gunnera tinctoria* and *Clethra arborea* [30, 31]. Most forest patches are dominated by *P. undulatum*, *Acacia melanoxylon* and *Eucalyptus globulus*, along with some remaining stands of former native forests including *Juniperus brevifolia, Morella faya* and *Laurus azorica.*

2.2 Field Sampling and Biomass Measurements

To develop *AGB* equations we used the data obtained from a survey undertaken between February and June 2010, which originated a large sample including 470 *P. undulatum* trees measured at 11 different exotic woodland stands in São Miguel Island. Tree density at each stand was estimated by using point-centered quarter method [32]. At each random point within the stand the distance to the center of the nearest tree was measured, at each of four quadrants. For each tree the following measures were taken: number of trunks at breast height (1.30 m), *NB*; diameter at breast height, *D* (using a diameter measuring tape); tree height, *H* and canopy height, *CH* (using a Vertex IV 360° and Transponder T3, Haglöf Sweden AB), and canopy diameter, *CD* (using a measuring tape). Tree *AGB* was measured by cutting the trees close to the base and weighting the resulting biomass using portable dynamometers.

2.3 Clustering of *P.* undulatum Stands

To summarize our data set and to determine if the *P. undulatum* stands could be grouped according to their general structure, we applied a Principal Component Analysis (PCA) using a general characterization of the sampled stands, namely tree density, mean *D*, mean *H* and mean *NB*. Since the amount of variance explained by the first two extracted factors was high (89%), and both variables and stands appeared as well discriminated in the corresponding plots, we applied a cluster analysis (Euclidean distance and UPGMA) followed by k-means cluster in order to define the number of groups and the stands belonging to each group, based on the dendrometric traits. Calculations were performed by using IBM SPSS Statistics, version 21.

2.4 Statistical Analyses - Biomass Allometric Equations

To estimate *AGB* of *P. undulatum* trees, we used several allometric equations, following [33, 34]. The most commonly used mathematical model is the allometric equation corresponding to the following power form:

$$Y = aX^b \tag{1}$$

where *a* and *b* are the scaling coefficients that vary with the variables under investigation, *Y* is the total biomass, and *X* a predictive variable(s) corresponding to tree dimension(s).

A total of 32 models were fitted to study the relationship between biomass and predictive variables. Firstly, we applied the allometric approach [33] by fitting linear and non linear models to the dataset (3 non-linear models, 2 weighted linear models, 2 linear models with variance model, 2 linear models with transforming variables, 2 simple linear models and 1 multiple linear model). Secondly, we tested 2 more linear models [34]. Finally, we tested 18 new multiple linear models.

For the linear models the original data were log-transformed and the least squares method was applied in order to estimate the parameters:

$$ln(Y) = a + b_1 ln(X) + \varepsilon \tag{2}$$

To select the type of model to be fitted and to specify both the form of the mean relation and the form of the error, we analyzed the scatter plots of *AGB* versus the predictive variables. In order to satisfy the prerequisite of linear regression, namely to equalize the variance over the entire range of biomass values, we transformed the data values using a natural logarithm [36]. The logarithmic equation is mathematically equivalent to Eq. (1), however they are not identical in a statistical sense [37]. Using the logarithmic form of Eq. (1), a systematic underestimation of the dependent variable Y is produced when converting the estimated $\log(Y)$ back to the original untransformed scale Y [36, 38]. We thus applied a correction factor (*CF*) when back transforming the calculations into biomass [36, 40, 41]. The *CF* is given by the following equation [36]:

$$CF = exp\left(\frac{(SEE \times 2.303)^2}{2}\right) \tag{3}$$

where *SEE* is the standard error of the estimate. A smaller *SEE* and *CF* indicate a higher model precision. Also, preliminary analyses included data exploration, namely histogram analysis, normality testing and residual plot examination, as recommended by the specialized literature [10].

Model comparison and selection were based on average deviation [42, 43], and on Akaike Information Criterion (*AIC*) [44]. The coefficient of determination (adjusted R^2) was used to determine the percentage of variance explained by each model. The predictive performance (goodness of fit) of the models was evaluated by calculating the root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{(\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}} \tag{4}$$

where X_{obs} is the observed value and X_{model} is the predicted value at i. We also calculated the mean error (difference between predicted an observed values) as a percentage of the estimated mean, here after designated as mean relative error (*MRE*).

After calculating the allometric equation, scatter plots were used to see whether the relationship between independent and dependent variables was linear. Besides the statistics above, we evaluated model performance by visual inspection. Therefore, scatter plots of residuals against the estimated values were analyzed to check the presence of the assumptions of linear regression, namely homoscedasticity. Model selection proceeded sequentially as follows: models with significant regression parameters; models with lower *AIC*; models with lower error (*RMSE*, *MRE*); models with higher R^2.

Statistical analyses (regressions and tests) were all performed within the R environment (R Development Core Team, 2014).

3 Results

3.1 Characterization of *P. undulatum* Stands

The variability in the diameter and height ranges between sites was large, with maximum heights varying between 4.00 and 16.70 m and maximum diameters at breast height ranging from 18.20 to 75.80 cm (sums of diameters for all the trunks at breast height per tree).This variability is indicative of some of the variation in site-specific diameter-to-height allometries. On the other hand, the variability in the number of branches between sites was quite large, with maximum values ranging from 7 to 81 (Table 1).

Based on tree density, mean *D*, mean *H*, mean *CH* and mean *NB*, it was possible to cluster the eleven *P. undulatum* stands into five groups. As stated in the Methods section, the first two components of the PCA explained 89% of the variance. Thus, it was possible to group the different stands according to the four variables used, as shown in the dendrogram (Fig. 1).

Analysis of the results obtained both with cluster analysis and k-means cluster, suggested the occurrence of five groups of stands.

Pittosporum undulatum stands from Group 1 show relatively high number of branches *NB* (53–81) but not the smallest *D* (58.30–66.60 cm) (Table 1, Fig. 2); stands from Group 2 are positioned at a more intermediate position; stands from Group 3 correspond to very tall trees *H* (11.00–16.70 m) with high *D* (63.00–75.80 cm) and low tree density (1725–3483 trees ha^{-1}); stands from Group 4 correspond to small trees *H* (4.00–5.50 m) with a high *NB* (37–54) and to a high tree density (15785–30121 trees ha^{-1}); the stand in Group 5 seems to be similar to those in group 4 but with a lower level of branching ($NB = 8$).

3.2 Allometric Equations

Several models performed poorly, namely three non-linear models(R^2 < 0.71, AIC$>$ 51, 4036), two weighted linear models (R^2 < 0.67, AIC$>$ 4547), two linear models with variance model (R^2 < 0.56, AIC$>$ 5144), and two linear models with transforming variables (R^2 < 0.75, AIC$>$ 5009). We obtained the best 23 models for the general data set (Table 2) from which the best seven models were selected (Table 3). For the general data set derived models (Table 2), the mean relative error was consistently above 20%. Among the best seven models, we found equations based on *BA* and *H* (Table 3, Fig. 3).

It should be stressed that in many cases models presented a high value of R^2 (Table 2), but in some cases the graphical exploration showed some deformed plots, with increasingly unrealistic extrapolations outside the range of the data, typical of model over-parameterization [33]. Also, some models showed an apparent good fit but included regression coefficients that were not significant. This complementary

Table 1 Allometric properties of 11 stands of *Pittosporum undulatum* in São Miguel Island, Azores: N = number of trees; Parameters (min. minimum, max. maximum, mean and standart deviation), Basal area (*BA*), Diameter at breast height (*D*), Maximum height (*H*), Canopy diameter (*CD*), Canopy height (*CH*), Number of branches (*NB*) and Biovolume (*BV*)

Stand	N	Density (ha^{-1})	Parameter	BA (cm^2)	D (cm)	H (m)	CD (m)	CH (m)	NB	BV (m^3)
1	37	3483	Min./Max.	2.01/856.23	1.60/75.80	1.90/16.70	0.58/6.49	0.30/10.80	1/7	2.27/899.35
			Mean	174.24	17.47	9.84	2.38	3.86	1.78	159.93
			Std.Dev.	183.00	15.22	3.44	1.31	2.55	1.49	213.54
2	38	5423	Min./Max.	0.10/147.30	0.76/25.75	0.90/6.60	0.50/6.70	0.60/4.10	1/8	0.09/408.77
			Mean	32.31	7.84	4.57	2.31	1.99	2.11	65.26
			Std.Dev.	36.74	5.16	1.27	1.45	0.93	1.78	81.55
3	37	11970	Min./Max.	0.38/143.29	0.70/27.60	1.17/4.20	0.43/2.98	0.15/2.60	1/7	0.32/44.47
			Mean	20.42	6.08	2.80	1.13	1.48	1.95	6.98
			Std.Dev.	26.15	5.16	0.78	0.51	0.62	1.27	8.57
4	38	15785	Min./Max.	2.33/127.99	3.10/50.90	1.62/4.00	0.66/3.47	0.27/2.70	1/54	1.37/55.58
			Mean	30.09	14.03	2.91	1.61	1.49	8.92	13.04
			Std.Dev.	24.24	9.67	0.62	0.52	0.66	9.40	9.77
5	64	1725	Min./Max.	0.64/1264.86	0.90/66.30	1.46/16.20	0.58/7.78	0.36/7.30	1/7	0.41/1777.39
			Mean	199.55	16.96	8.13	3.28	3.20	1.92	275.99
			Std.Dev.	253.43	13.56	3.53	1.51	1.71	1.51	368.96
6	31	2782	Min./Max.	6.60/346.00	2.90/63.00	3.90/11.00	0.87/5.53	0.10/6.50	1/27	5.94/728.46
			Mean	137.47	22.77	8.03	3.26	1.37	5.77	262.53
			Std.Dev.	85.09	12.89	1.55	1.14	1.63	5.61	182.21

(continued)

Table 1 (continued)

Stand	N	Density (ha^{-1})	Parameter	BA (cm^2)	D (cm)	H (m)	CD (m)	CH (m)	NB	BV (m^3)
7	70	4796	Min./Max.	1.33/140.99	1.3/66.60	1.87/6.22	0.51/3.49	0.10/2.89	1/81	2.15/103.01
			Mean	46.31	25.05	3.93	1.88	1.40	16.96	31.49
			Std.Dev.	30.11	15.96	1.30	0.57	0.71	16.45	24.12
8	39	10094	Min./Max.	2.01/148.8	1.60/58.30	2.56/6.10	1.12/3.33	0.27/2.30	1/53	2.13/156.22
			Mean	53.22	27.46	4.71	2.32	1.01	15.69	68.70
			Std.Dev.	32.50	13.97	0.80	0.61	0.56	11.42	38.21
9	40	37995	Min./Max.	0.71/51.26	1.20/18.20	1.82/7.20	0.44/2.45	0.24/4.10	1/8	0.82/58.18
			Mean	11.49	5.62	4.56	1.44	2.08	2.83	18.78
			Std.Dev.	10.23	4.11	1.11	0.45	1.07	1.99	14.51
10	39	13231	Min./Max.	0.28/118.6	0.60/32.00	0.80/12.00	0.76/4.50	0.28/8.10	1/8	1.94/383.20
			Mean	25.13	6.43	5.58	2.07	2.15	1.79	62.74
			Std.Dev.	24.10	5.06	2.05	0.77	1.70	1.5	80.12
11	38	30121	Min./Max.	0.39/40.23	0.80/34.60	1.16/5.50	0.54/3.19	0.35/2.50	1/37	0.76/74.35
			Mean	9.93	9.75	3.02	1.66	1.22	10.71	20.75
			Std.Dev.	10.07	8.37	0.94	0.62	0.52	8.90	19.83

Note $BA = (D^2 \times \pi/4)$; $BV = ((H - CH) \times (CD/2)^2 \times \pi)$; *D* and *BA* result from the sum of all branches per tree. Stands:1- Pico das Camarinhas - caldera; 2 - Pico das Camarinhas - east slope; 3 - Pico das Camarinhas - west slope; 4 - Pico das Camarinhas - top; 5 - Lagoa de Santiago; 6 - São Roque; 7 - Old landfill; 8 - Picos de Lima; 9 - Pinhal da Paz; 10 - Furnas; 11 - Cabouco

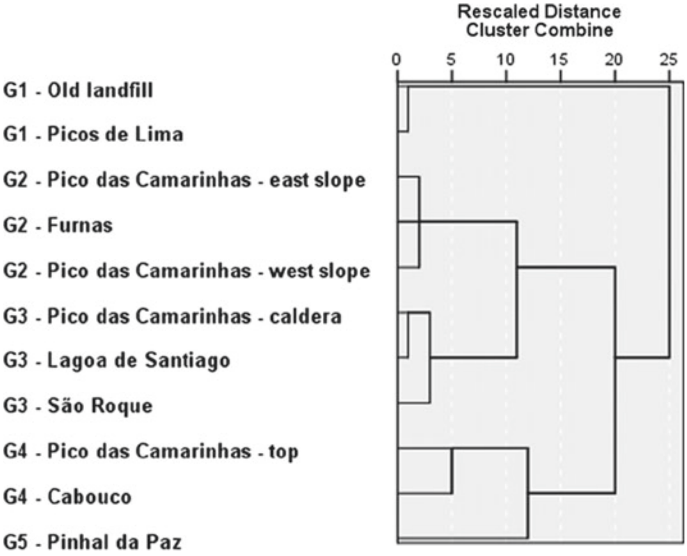


Fig. 1 Results of a cluster analysis (Euclidean Distance and UPGMA) applied to 11 *Pittosporum undulatum* stands in São Miguel Island. The variables used to group the stands were mean tree density, mean *D*, mean *H*, and mean *NB*

Fig. 2 Principal Component Analysis plot of the 11 *Pittosporum undulatum* stands in São Miguel Island. See Fig. 1 for groups description

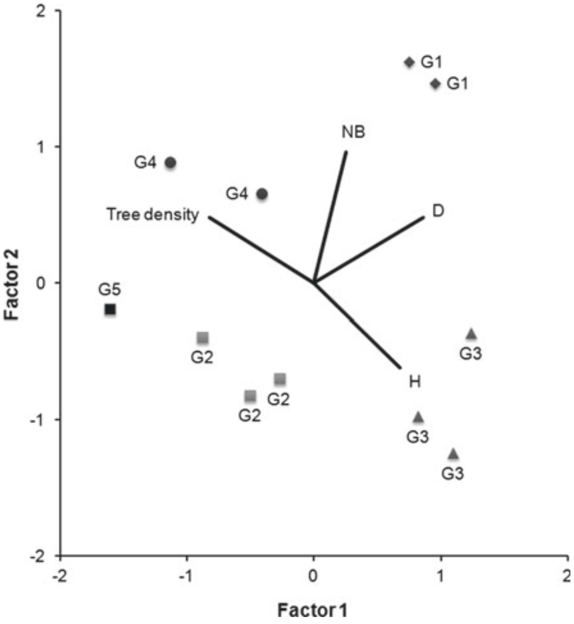


Table 2 Allometric equations tested for 470 *Pittosporum undulatum* trees in São Miguel Island, Azores. B: Biomass. Allometric models: * [33]; ** [34]; ***Current study. Model equation; adjusted determination coefficient (Adj R^2); Akaike Information Criterion (AIC); Root Mean Square Error (RMSE); Mean Relative Error (MRE)

Model	Regression Model	Adj R^2	AIC	RMSE	MRE
1	ln(B)= a + $b_1 ln(D) + \varepsilon$ *	0.50	1515.38	81.52	1.10
2	ln(B)= a + $b_1 ln(D) + b_2 ln(H) + \varepsilon$ *	0.81	1047.00	47.75	0.27
3	ln(B)= a + $b_1 ln(D) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 ln(BV) + \varepsilon$ **	0.84	975.46	44.75	0.24
4	ln(B)= a + $b_1 ln(BA) + \varepsilon$**	0.78	1132.26	33.47	0.36
5	ln(B)= a + $b_1 ln(BA) + b_2 ln(H) + \varepsilon$ **	0.83	1012.99	32.62	0.26
6 #	ln(B)= a + $b_1 ln(BA) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 ln(BV) + \varepsilon$ **	0.87	889.67	32.70	0.20
7 #	ln(B)= a + $b_1 ln(D) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 ln(BV) + b_6 ln(NB) + \varepsilon$ **	0.86	912.75	34.55	0.21
8 #	ln(B)= a + $b_1 ln(BA) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 ln(BV) + b_6 ln(NB) + \varepsilon$ **	0.87	889.99	33.22	0.20
9	ln(B)=$a + b_1 ln(D^2 H) + \varepsilon$ *	0.66	1329.02	66.92	0.64
10	ln(B)= $a + b_1 ln(D^2 H) + b_2 ln(H) + \varepsilon$***	0.81	1047.49	47.75	0.27
11	ln(B)= $a + b_1 ln(D^2 H) + b_2 ln(H) + b_3 ln(CH) + \varepsilon$***	0.82	1045.30	48.57	0.27
12	ln(B)= $a + b_1 ln(D^2 H) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 (BV) + \varepsilon$***	0.84	975.47	44.75	0.24
13	ln(B)= $a + b_1 ln(BA^2 H) + \varepsilon$***	0.81	1055.10	29.21	0.29
14	ln(B)= $a + b_1 ln(BA^2 H) + b_2 ln(H) + \varepsilon$ ***	0.83	1012.99	32.62	0.26
15 #	ln(B)= a + $b_1 ln(BA^2 H) + b_2 ln(H) + b_3 ln(CH) + \varepsilon$ ***	0.83	1014.39	32.42	0.25
16 #	ln(B)= $a + b_1 ln(BA^2 H) + b_2 ln(H) + b_3 ln(CH) + b_4 ln(CD) + b_5 ln(BV) + b_6 ln(NB) + \varepsilon$***	0.87	889.99	33.22	0.20
17 #	ln(B)= $a + b_1 ln(BA^2 H) + b_2 ln(H) + b_3 ln(NB) + \varepsilon$***	0.83	1011.65	34.32	0.25
18	ln(B)= $a + b_1 ln(BA^2 H) + b_2 ln(NB) + \varepsilon$ ***	0.81	1050.35	28.14	0.29
19 #	ln(B)= $a + b_1 ln(BA^2 H) + b_2 ln(H) + b_3 ln(BV) + \varepsilon$***	0.86	921.83	29.41	0.21

(continued)

Table 2 (continued)

Model	Regression Model	Adj R^2	AIC	RMSE	MRE
20	ln(B)= $a + b_1 ln(BA^2H) + b_2 ln(H) + b_3 ln(BA) + \varepsilon$***	0.83	1012.99	32.62	0.26
21 #	ln(B)= $a + b_1 ln(BA^2H) + b_2 ln(H^2) + b_3 ln(NB^2) + \varepsilon$ ***	0.83	1011.65	34.32	0.25
22	ln(B)= $a + b_1 ln(BA^2H) + b_2 ln(H^2) + \varepsilon$***	0.83	1012.99	32.62	0.26
23	ln(B)= $a + b_1 ln(BA^2H) + b_2 ln(D^2H) + b_3 ln(H^2) + b_4 ln(NB^2) + \varepsilon$***	0.84	967.04	33.13	0.23

Note # Equations that included non-significant regression coefficients ($P > 0.05$)

analysis affected model selection. Model 23 was the best of the models with all the regression coefficients significant. It included *D*, *BA*, *H* and *NB*. Models 3 and 12 followed, which included data about tree canopy, and that were almost equivalent. Models 5, 14, 20 and 22 were all very similar, including data on *BA* and *H*, with somewhat higher *AIC* but lower *RMSE* values than the previous group.

In the case of the general data set (Table 3), the correction factor values determined for the various log-transformed allometric equations were always above one. This might imply some degree of error in the general predictions derived from those models, since only values of *CF* close to one indicate a highly significant relationship of the dependent and independent variables [45].

Regarding the analysis of the individual data sets for each of the five groups of *P. undulatum* stands, the best models, selected among the 23 allometric equations tested, were obtained when using *D*, *BA* and *H* as predictive variables (Table 4, Fig. 4). Even when considering the models derived for each type of *P. undulatum* stand, the value of *CF* varied considerably. In case of groups 1, 3 and 5 the values of *CF* were very low, indicating that the downward bias of the equations was marginal [46], what is in agreement with a mean relative error of the prediction below 20% (Table 4). On the other hand, groups 2 and 4 showed an increase of the mean relative error of the prediction and on the values of the *CF*, when compared to the values obtained for the best models derived from the entire dataset.

In general, the residuals showed a typical distribution that is symmetric (Fig. 5). Also, most of the residuals follow a normal distribution, with the exception of a few more extreme values.

4 Discussion

Tree biomass can be estimated as a function of the diameter at breast height, or of the related basal area, of tree height and tree density at a given location. However, the contribution of these parameters to the estimation of above ground biomass differs according to the site, successional stage of the forest, disturbance level, species composition, among other factors [15, 16]. Therefore, the choice of an allometric model depends of its statistical properties as well as of the specific context for its

Table 3 Selected allometric equations tested for 470 *Pittosporum undulatum* trees found at eleven exotic woodland stands in São Miguel Island, Azores. Plot number (see Fig. 3); Regression model (see Table 2); Model parameters: Coefficient value, P value and Correction Factor (CF)

Plot Number	Regression Model*	Model parameters			
		Coefficient		P value	CF**
1	3	a	−1.38	<0.0001	1.26
		b_1	0.63	<0.0001	
		b_2	1.10	<0.0001	
		b_3	0.25	<0.0001	
		b_4	0.33	0.0255	
		b_5	0.37	0.0003	
2	5	a	−1.31	<0.0001	1.28
		b_1	0.72	<0.0001	
		b_2	1.05	<0.0001	
3	12	a	−1.38	<0.0001	1.26
		b_1	0.31	<0.0001	
		b_2	0.79	<0.0001	
		b_3	0.25	<0.0001	
		b_4	0.33	0.0255	
		b_5	0.37	0.0003	
4	14	a	−1.31	<0.0001	1.28
		b_1	0.36	<0.0001	
		b_2	0.69	<0.0001	
5	20	a	−1.31	<0.0001	1.28
		b_1	0.33	<0.0011	
		b_2	1.37	<0.0001	
		b_3	0.71	<0.0001	
6	22	a	−1.31	<0.0001	1.28
		b_1	0.36	<0.0001	
		b_2	0.34	<0.0001	
7	23	a	−1.64	<0.0001	1.25
		b_1	0.16	<0.0001	
		b_2	0.29	<0.0001	
		b_3	0.41	<0.0001	
		b_4	−0.15	<0.0001	

*Equation before application of Correction Factor (CF)

**Correction Factor to remove the bias of regression estimates after logarithmic transformed data

application. In the Azores Islands, this is one of the first examples of a study, based on a large sample, dedicated to one of the most abundant forest species, *P. undulatum*, the most important woody plant invader in the Azores, presently being aimed as a biomass resource for renewable fuel production [22].

Table 4 Allometric models applied to five groups of *Pittosporum undulatum* stands in São Miguel Island. Group number, plot number (see Fig. 4), regression model (see Table 2), model coefficients, P value, adjusted determination coefficient (Adj R^2), Root Mean Square Error (RMSE), Mean Relative Error (MRE), Akaike Information Criterion (AIC), Correction Factor (CF)

Group number	Plot number	Regression model*	Model parameter			Model performances				
			Coefficient		P value	Adj R^2	RMSE	MRE	AIC	CF**
1	1	13	*a*	−0.87	<0.0001	0.80	7.80	0.07	109.00	1.08
			b_1	0.43	<0.0001					
2	2	2	*a*	−2.26	<0.0001	0.72	25.79	0.41	310.20	1.54
			b_1	0.95	<0.0001					
			b_2	2.00	<0.0001					
3	3	13	*a*	−1.65	<0.0001	0.94	46.73	0.14	137.67	1.08
			b_1	0.51	<0.0001					
					<0.0001					
4	4	5	*a*	−1.37	<0.0001	0.66	6.69	0.29	173.04	1.30
			b_1	0.54	<0.0001					
			b_2	1.58	<0.0001					
5	5	14	*a*	−2.59	<0.0001	0.83	2.72	0.12	57.19	1.11
			b_1	1.56	<0.0001					
			b_2	0.33	0.0014					

*Equation before application of Correction Factor (CF)

**Correction Factor to remove the bias of regression estimates after logarithmic transformed data

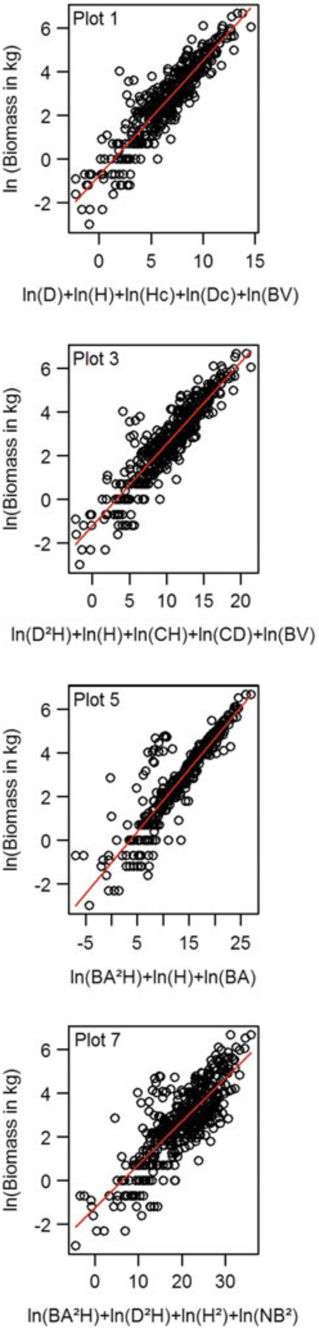
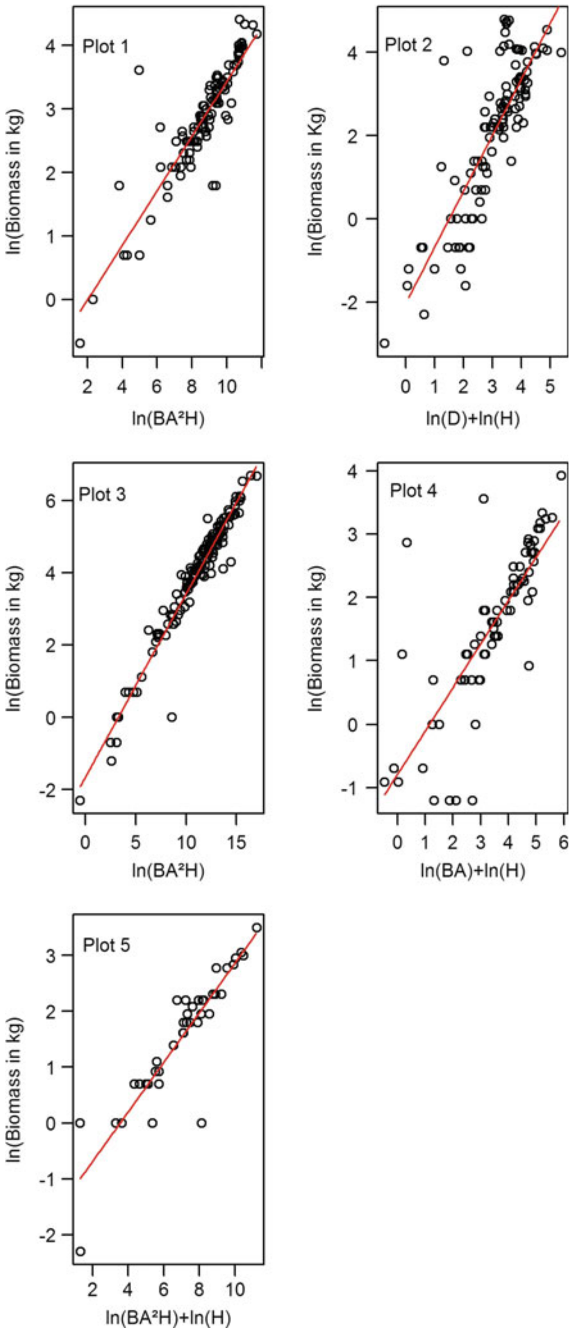


Fig. 3 Plots of the allometric equations used to predict *Pittosporum undulatum* biomass from dendrometric traits, based on a total sample of 470 trees. Plot number as defined in Table 3

Fig. 4 Plots of the allometric equations used to predict *Pittosporum undulatum* biomass from dendrometric traits, for each of five groups of stands. Plot number as defined in Table 4



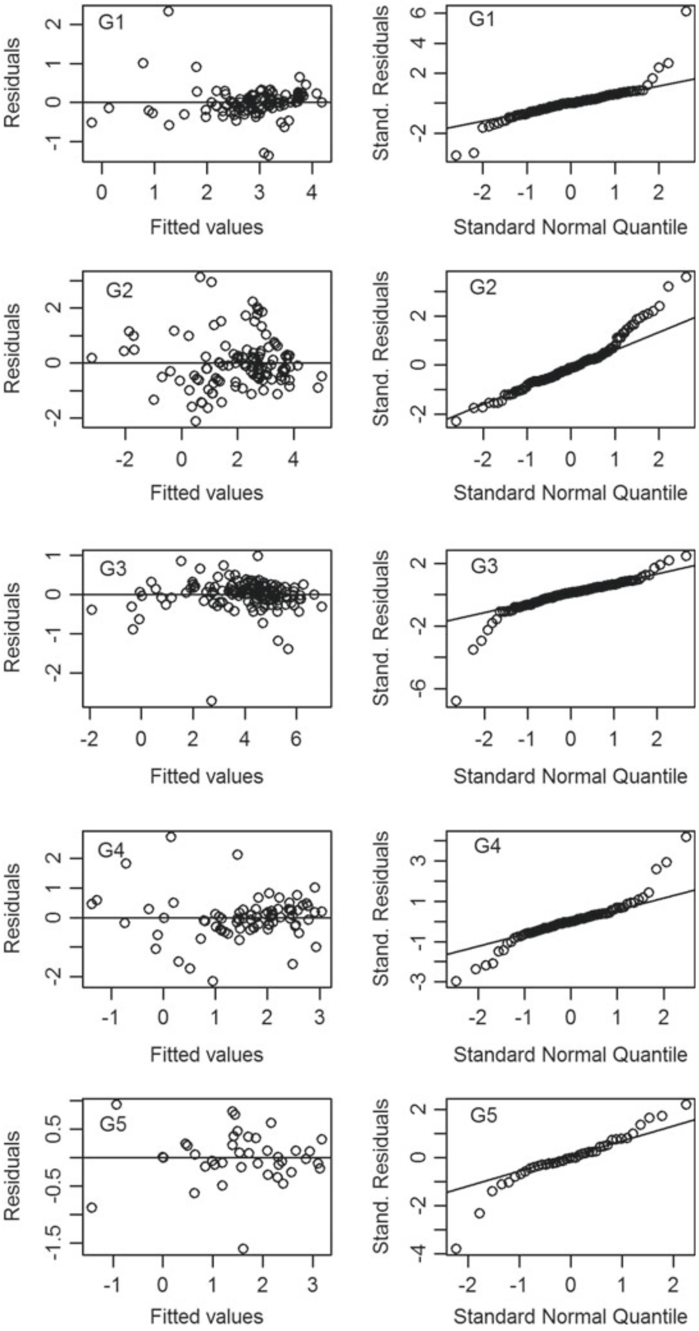


Fig. 5 Plots of residuals of the relationships between the logarithm of biomass and the predictor variables for the best model in each group of stands (G1-G5). *Left side* fitted values versus residuals. *Right side* normal quantile - quantile plot

According to our results, *AGB* was frequently explained by models that included *D*, *BA* and *H*. Diameter at breast height is the common and best predictor for biomass in allometric models because both variables are strongly correlated, namely in conifer equations [49, 50]; in addition, *D* can be easily measured in the field and is always available in forest inventories data [7, 51]. However, our results showed that *BA* frequently improved model adjustment when compared to allometric equations including *D*. A strong relationship of biomass with basal area was found in many studies [52, 53]. However, improvement in the regression was reported by using basal area [15, 52] and height [54]. Similarly, in our study the interaction term $BA \times H$ was more appropriate than $D \times H$ to estimate biomass. In the same sense, it has been considered [51] that the interaction term $D^2 \times H$ is more appropriate than $D \times H$ to estimate biomass, since the latter is a product of volume ($\pi/4 \times D^2 \times H \times$ form factor).

Models that incorporate *H* and *CH* usually give good fits [55] but in many cases these measurements are difficult to carry out. Most height measurements are sources of pronounced error for large trees. Canopy height is particularly difficult to measure in trees with irregular canopy structure such as *P. undulatum*. Five sources of uncertainty may affect the precision of tree height measurements [56]: offset between measured distance and crown-top position, tree-top occlusion, ground slope, obstacles for distance measurements, and clinometer operator error. Although the use of canopy diameter has been reported to estimate tree biomass [57], in our study this type of model was found to be among the seven best, but originated a somewhat high level for *RMSE*. Therefore, we don't recommend the use of those models. Besides avoiding further measuring errors, this will also be an advantage in practical terms, since the estimation of *AGB* will generally only demand the record of *D*, *H* and eventually *NB*, avoiding the more time consuming measurement of other canopy dimensions.

Several studies also include wood density as predictor variable for estimating above-ground biomass [59, 60]. However, the role of wood density is more prominent for mixed species, that can differ greatly in tree architecture and wood density, than for species-specific allometric equations [61–63].

Our results highlight that the decision to log-transform raw data in allometry is more than one of statistical convenience. It has been argued that fitting linear models on log transformed data leads to results that are‘ biased and misleading‘ because such models operate in geometric rather than arithmetic space, and that analyses should be performed on the original scale [47]. However, other researchers [48] note that many allometric characteristics of organisms are‘ multiplicative by nature‘ and thus fitting models to log-transformed data is perfectly acceptable because accounting for proportional rather than absolute variation is most important.

As stated above, differences in the contribution of the different parameters to the estimation of *AGB* are expected, due to changes in tree structure that occur along time and space. For instance, trees located at stands with different topographical characteristics (e.g. slope, degree of exposure, aspect) will in turn show differences in growth rate, shape, and architecture. For this reason, we analyzed model fit for the entire data set and also separately, according to the groups of stands obtained from a multivariate analysis. We found considerable variation in stand structure, including from dense stands, dominated by relatively small multi-branched trees, up to low density stands dominated by very tall trees with large *D*. Nevertheless, it was possible to group the stands according to tree density, *D*, *H* and *NB*. In the same sense, several differences in biomass allocation in stone pine have been found [58], resulting from different stand characteristics, emphasizing the importance of stand-dependent factors for adjusting regional or national biomass calculations.This was evident in our research since when modelling tree biomass per group of homogenous stands, only *D*, *BA* and *H* were needed, while when modeling the global data set which included heterogeneous stands *N* was also needed. The latter possibly accounted for structural differences among trees from different groups of stands. Therefore, implementing allometric equations beyond the specific site and the diameter range for which they were developed could affect the accuracy of biomass estimates.

The present study suggests that if data on *BA* and *H* is available, a relatively simple equation could be used for the estimation of *P. undulatum AGB*. However, researchers and managers should be aware that the high variability in the structure of *P. undulatum* stands will always originate some level of error when estimating biomass resources. Our results underscore the importance of stand-dependent factors and the need for calibrating biomass estimates when using allometric equations. For this reason, work is under way to further increase the information available about *P. undulatum* stands in the Azores islands, namely at São Miguel, Terceira and Pico islands.

Although this is a case study for a single region, findings based on this study could also be relevant to forests in other regions. Finally, we hope that this study, in conjunction with research devoted to tree age determination and species distribution modeling, will support an effective management of *P. undulatum* in the Azores, allowing the effective containment of the invasion while originating economic return and social benefits from the extracted biomass.

Acknowledgements The first author would like to acknowledge to NATURALREASON, Lda, a company devoted to Biofuels Production, for supporting her research grant and for its financial contribution to the ongoing research project.This work was also supported by Project Woody Biomass (Azorean Governement and MIT Portugal) and by Direção Regional dos Recursos Florestais (Azores). We also thank an anonymous reviewer for useful suggestions. The final writing of this paper was supported by project FOREST-ECO2—Towards an Ecological and economic valorization of the Azorean Forest—ACORES-01-0145-FEDER-000014—Azores 2020 PO, 2016–2019.

References

1. Henry, M., Picard, N., Trotta, C., Manlay, R.J., Valentini, R., Bernoux, M., Saint-André, L.: Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. Silva Fennica **45**(3B), 477–569 (2011)
2. Canga, E., Dieguez-Aranda, I., Afif-Khouris, E., and Camara-Oregon, A.: Above-ground biomass equations for *Pinus radiata* D. Don in Asturias. Instituto Nacional de Investigacin y Tecnologa Agraria y Alimentaria (INIA). For. Syst. **22**(3), 408–415 (2013)

3. Zianis, D., Xanthopoulos, G., Kalabokidis, K., Kazakis, G., Ghosn, D., Roussou, O.: Allometric equations for aboveground biomass estimation by size class for *Pinus brutia* Ten. Trees growing in North and South Aegean Islands, Greece. Eur. J. For. Res. **130**, 145–160 (2011)
4. Palm, W.J.: System Dynamics, 2nd edn. McGraw Hill, New York (2010). p. 225
5. Zianis, D., Mencuccini, M.: On simplifying allometric analyses of forest biomass. For. Ecol. Manag. **187**, 311–332 (2004)
6. Moore, J.R.: Allometric equations to predict the total aboveground biomass of radiata pine trees. Ann. For. Sci. **67**(8), 806 (2010)
7. Zianis, D., Mencuccine, M.: Aboveground biomass relationships for beech (*Fagus moesiaca* Cz.) trees in Vermio Mountain, Northern Greece, and generalised equations for Fagus sp. Ann. For. Sci. **60**, 439–448 (2003)
8. Addo-Fordjour, P., Rahmad, Z.B., Shahrul, A.M.S.: Effects of human disturbance on liana community diversity and structure in a tropical rainforest, Malaysia: implication for conservation. J. Plant Ecol. **5**(4), 391–399 (2012)
9. Arevalo, C.B.M., Volk, T.A., Bevilacqua, E., Abrahamson, L.: Development and validation of aboveground biomass estimations for four Salix clones in central New York. Biomass and Bioenergy **31**, 1–12 (2007)
10. Picard, N., Henry, M., Mortier, F., Trotta, C., Saint-André, L.: Using Bayesian model averaging to predict tree aboveground biomass. For. Sci. **58**(1), 15–23 (2012)
11. Ebuy, J., Lokomb, D.J.P., Ponette, Q., Sonwa, D., Picard, N.: Biomass equation for predicting tree aboveground biomass at Yangambi. DRC. J. Tropical For. Sci. **23**(2), 125–132 (2011)
12. Genet, A., Wernsdörfer, H., Jonard, M., Pretzsch, H., Rauch, M., Ponette, Q., Nys, C., Legout, A., Ranger, J., Vallet, P., Saint-André, L.: Ontogeny partly explains the apparent heterogeneity of published biomass equations for *Fagus sylvatica* in central Europe. For. Ecol. Manag. **261**(7), 1188–1202 (2011)
13. Van, B., Ransijn, M.J., Craven, D., Bongers, F., Hall, J.S.: Estimating carbon stock in secondary forests: decisions and uncertainties associated with allometric biomass models. For. Ecol. Manag. **262**(8), 1648–1657 (2011)
14. Ketterings, Q.M., Coe, R., Van, M.N., Russell, A.E.: Impact of spatial variability of Y. Ambagau and C.A. Palm, 2001. Reducing tropical forest structure on radar estimation of uncertainty in the use of allometric biomass aboveground biomass, Remote Sensing of equations for predicting above-ground tree biomass. Environment **115**, 2836–2849 (2011)
15. Brunig, E.F.: Structure and growth. In: Golley, F.B. (ed.) Ecosystems of the World 14A, Tropical Rain Forest Ecosystems: Structure and Function, pp. 49–75. Elsevier Scientific publication, New York (1983)
16. Whitmore, T.C.: Tropical Rainforests of the Far East, pp. 112–113. Oxford University Press, London (1984)
17. Torres, A.B., Lovett, J.C.: Using basal area to estimate aboveground carbon stocks in forests: La Primavera Biosphere‘s Reserve. Mexico. For. **86**, 267–281 (2013)
18. Burrows, W.H., Hoffmann, M.B., Compton, J.F., Back, P.V., Tait, L.J.: Allometric relationships and community biomass estimates for some dominant eucalypts in Central Queensland woodlands. Aust. J. Bot. **48**, 707–714 (2000)
19. Sarmiento, G., Pinillos, M., Garay, I.: Biomass variability in tropical American lowland rainforests. Ecotropicos **18**, 1–20 (2005)
20. Trelease, W.: Botanical observationson the Azores. Ann. Rep. Missouri Botanical Gard. **8**, 77–220 (1897)
21. Silva, L., Ojeda-Land, E., Rodríguez-Luengo, J.L.: Invasive Terrestrial Flora and Fauna of Macaronesia. Top 100 in Azores, Madeira and Canaries. ARENA, pp. 546. Ponta Delgada (2008)
22. Lourenço, P., Medeiros, V., Gil, A., Silva, L.: Distribution, habitat and biomass of *Pittosporum undulatum,* the most important woody plant invader in the Azores Archipelago. For. Ecol. Manag. **262**(2), 178–187 (2011)
23. Costa, H., Aranda, S., Lourenço, P., Medeiros, V., Azevedo, E.B., Silva, L.: Predicting successful replacement of forest invaders by native species using species distribution models: The

case of *Pittosporum undulatum* and *Morella faya* in the Azores. For. Ecol. Manag. **279**, 90–96 (2012)

24. Costa, H., Medeiros, V., Azevedo, E.B., Silva, L.: Evaluating ecological-niche factor analysis as a modeling tool for environmental weed management in island systems. Weed Res. **53**, 221–230 (2013)
25. Forjaz, V.H., Tavares, J., Brito, E., Rodrigues, M., Gonçalves, J., Nunes, J., Santos, R., Barreiros, J., Gallagher, L., Silva, P., Barcelos, P., França, Z., Dentinho, T., Silva, V., Serpa, M.C., Magalhães, L.: The Azores Basic Atlas. Observatório Vulcanológico e Geotérmico dos Açores. Ponta Delgada, Azores (2004)
26. Ferreira, D.B.: Contribution à l' étude des ventes et de l'humidité dans les iles centrales de l'archipel des Açores. Centro Estudos Geográficos, Lisboa (1980)
27. Agostinho, J.: Clima dos Açores, parte 1. Açoreana **2**, 35–65 (1938)
28. Carine, M.A., Schaefer, H.: The Azorean diversity enigma: why are there so few Azorean endemic flowering plants and why are they so widespread? J. Biogeogr. **37**, 77–89 (2010)
29. Dias, E.: Vegetação natural dos Açores. Ph.D. thesis. Universidade dos Açores, Angra do Heroísmo, pp. 302 (1996)
30. Silva, L.: Plantas Vasculares Invasoras no Arquipélago dos Açores. Caracterização Geral e Estudo de um Caso: *Clethra arborea* Aiton, Cletheraceae. Tese de Doutoramento, p. 541. Universidade dos Açores, Ponta Delgada (2001)
31. Silva, L., Smith, C.A.: Quantitative approach to the study of non-indigenous plants: an example from the Azores Archipelago. Biodivers. Conserv. **15**, 1661–1679 (2006)
32. Mitchell, K.: Quantitative Analysis by the Point Centered Quarter Method. Department of Mathematics and Computer Science. Geneva, NY (2007)
33. Picard N., Saint-André L., Henry M.: Manual for building tree volume and biomass allometric equations:from field measurement to prediction. Food and Agricultural Organization of the United Nations, Rome, and Centre de Coopration Internationale en Recherche Agronomique pour le Dveloppement, pp 215. Montpellier (2012)
34. Murali, K.S., Bhat, D.M., Ravindranath, N.H.: Biomass estimation equations for tropical deciduous and evergreen forests. Int. J. Agric. Res. Gov. Ecol. **4** (1), 81–92 (2005)
35. Sokal, R.R., Rohlf, F.J.: Biometry. The Principles and Practice of Statistic in Biological Research, 3rd edn. W.H. Freeman and Co., New York (1995). pp. 856
36. Sprugel, D.C.: Correcting for bias in log-transformed allometric equations. Ecology **64**(1), 209–210 (1983)
37. Zar, I.H.: Calculation and miscalculation of the allometric equation as a model in biological data. Biosciences **18**, 1118–1120 (1968)
38. Finney, D.I.: On the distribution of a variate whose logarithm is normally distributed. I. R. Stat. Sci. Ser. B. **7**, 155–161 (1941)
39. Mountford, M.D., Bunce, R.G.H.: Regression sampling with allometrically related variables, with particular reference to production studies. Forestry **46**, 203–212 (1973)
40. Sah, J.P., Ross, M.S., Koptur, S., Snyder, J.R.: Estimating aboveground biomass of broadleaved Woody plants in the understory of Florida Keys pine forest. For. Ecol. Manag. **203**(1–3), 319–329 (2004)
41. Son, Y., Hwang, J.W., Kim, Z.S., Lee, W.K., Kim, J.S.: Allometry and biomass of Korean pine *(Pinus koraiensis)* in Central Korea. Bioresour. Technol. **78**, 251–255 (2001)
42. Cairns, M.A., Olmsted, I., Granados, J., Argaez, J.: Composition and aboveground tree biomass of a dry semi-evergreen forest on Mexico's Yucatan Peninsula. For. Ecol. Manag. **186**, 125–132 (2003)
43. Nelson, B.W., Mesquita, R., Pereira, J.L.G., de Souza, S.G.A., Batista, G.T., Couta, L.B.: Allometric regressions for improved estimate of secondary forest biomass in the Central Amazon. For. Ecol. Manag. **117**, 149–167 (1999)
44. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach, 2nd edn. Springer Science + Business Media, Inc., New York (2002). pp. 488

45. Makungwa, S.D., Chittpck, A., Skole, D.L., Kanyama-Phiri, G.Y.: Allometry for biomass estimation in Jatropha trees planted as boundary hedge in farmers? Fields. For. **4**, 218–233 (2013)
46. Addo-Fordjour, P., Rahmad, Z.B.: Mixed species allometric models for estimating aboveground Liana Biomass in tropical primary and secondary forests, Ghana, pp. 1–9. Hindawi Publishing Corporation (2013)
47. Packard, G.C., Boardman, T.J.: Model selection and logarithmic transformation in allometric analysis. Physiol. Biochem. Zool. **81**, 496–507 (2008)
48. Kerkhoff, A.J., Enquist, B.J.: Multiplicative by nature: why logarithmic transformation is necessary in allometry. J. Theor. Biol. **257**, 519–521 (2009)
49. Jenkins, J.C., Chojancky D.C., Heath L.S., Birdsey R.: Comprehensive database of diameter-based biomass regressions for North american tree species. US For. Serv. Gren. Tech. Rep. NE-319.US For.Ser.Northeast.Res.Stn.,Newtown Square,Pp.45 (2004)
50. Tinker, D.G., Stakes, K.A.: Allometric equation development, biomass and aboveground productivity in ponderosa pine forests Black Hill, Wyoming. West. J. Appl. For. **25**, 112–119 (2010)
51. Vahedi, A.A., Mataji, A., Babayi-Kafaki, S., Eshaghi-Rad, J., Hodjati, S.M., Djomo, A.: Allometric equations for predicting aboveground biomass of beech-hornbeam stands in the Hyrcanian forests of Iran. J. For. Sci. **60**(6), 236–247 (2014)
52. Cannell, M.G.R.: Woody biomass of forest stands. For. Ecol. Manag. **8**, 299–312 (1984)
53. Rai, S.N., Proctor, J.: Ecological studies on four forests in Karnataka, India. I. Environment, structure, floristics and biomass. J. Ecol. **74**, 439–454 (1986)
54. O'Neill, R.V., De Angelis, D.L.: Comparative productivity and biomass relations of forest ecosystems? In: O'Neil, R.V., DeAngelis (eds.) Dynamic Properties of Forest Ecosystems, pp. 411–449. Cambridge University Press (1988)
55. Aráujo, T.M., Higuchi, N., Carvalho, J.A.: Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Par. Brazil. For. Ecol. Manag. **117**, 43–52 (1999)
56. Hunter, M.O., Keller, M., Victoria, D., Morton, D.C.: Tree height and tropical forest biomass estimation. Biogeosciences **10**, 8385–8399 (2013)
57. Mcginnis, T.W., Keeley, J.E.: Post-fore treatments impacts one fine fuels in westside Sierra Nevada forests. USGS Western Ecological Research Center (2010)
58. Cutini, A., Chianucci, F., and Manetti, M.C.: Allometric relationships for volume and biomass for stone pine *(Pinus pinea L.)* in Italian coastal stands. iForest, **6**, 331–337 (2013)
59. Cai, S., Kang, X., Zhang, L.: Allometric models for aboveground biomass of ten tree species in Northeast China. Ann. For. Res. **56**(1), 105–122 (2013)
60. Junior, L.N.R., Engel, V.L., Parrotta, J.A., Melo, A.C.G., Ré, D.S.: Allometric equations for estimating tree biomass in restored mixed-species Atlantic Forest stands.Biota. Neotropica **14**(2), 1–9 (2014)
61. Basuki, T.M., Van, L.P.E., Skidmore, A.K., Hussin, Y.A.: Allometric equations for estimating the above-ground biomass in tropical lowland Dipterocarp forests. For. Ecol. Manag. **257**, 1684–1694 (2009)
62. Djomo, A.N., Adamou, I., Joachim, S., Gode, G.: Allommetric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. For. Ecol. Manag. **260**, 1873–1885 (2010)
63. Alvarez, E., Duque, A., Saldarriaga, J., Cabrera, K., Salas, G.D.L., Valle, L.D., Lema, A., Moreno, F., Orrego, S., Rodriguez, L.: Tree above-ground biomass allometries for carbon stocks estimation in the natural forests of Colombia. For. Ecol. Manag. **267**, 297–308 (2012)

Alternating Hadamard Series and Some Theorems on Strongly Regular Graphs

Luís António de Almeida Vieira and Vasco Moço Mano

Abstract In this paper we consider a strongly regular graph, G, whose adjacency matrix A has three distinct eigenvalues, and a particular real three dimensional Euclidean Jordan subalgebra with rank three of the Euclidean algebra of real symmetric matrices of order n, with the product and the inner product being the Jordan product and the usual trace of matrices, respectively. Next, we compute the unique Jordan frame $\mathscr{B}$ associated to A and we consider particular alternating Hadamard series constructed from the idempotents of $\mathscr{B}$. Finally, by the analysis of the spectra of the sums of these alternating Hadamard series we deduce some theorems over the parameters of a strongly regular graph.

Keywords Graphs and linear algebra · Algebraic combinatorics · Graph theory

1 Introduction

In this chapter we establish some inequalities on the parameters of a strongly regular graph like we have done in the papers [11–13, 16] but recurring to alternating Hadamard series of a matrix.

Euclidean Jordan algebras have a lot of applications to many branches of mathematics, for instance in statistics (see [14]), interior point methods (see [3, 6, 7]

L.A. de Almeida Vieira (✉)
CMUP-Center of Research of Mathematics of University of Porto,
Department of Mathematics of Faculty of Sciences of University of Porto,
Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
e-mail: lvieira@fe.up.pt

V.M. Mano
Department of Mathematics, University of Porto, Porto, Portugal
e-mail: vascomocomano@gmail.com

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_23

and combinatorics (see [2, 11–13, 16]). In this paper we apply the theory of Euclidean Jordan algebras to strongly regular graphs to present some theorems over their parameters.

This paper is organized as follows. In Sects. 2 and 3, we present some preliminary definitions and results on Euclidean Jordan algebras and strongly regular graphs, respectively, that will be used throughout this paper. In Sect. 4 we associate a particular real Euclidean Jordan algebra to the adjacency matrix of a strongly regular graph, A, and we consider the unique Jordan frame, $\mathscr{B}$, associated to A. Next, we establish some theorems over the parameters and the spectra of a strongly regular graph by the analysis of the spectra of an alternating Hadamard series of an element of $\mathscr{B}$. Finally, in Sect. 5, we present some experimental results.

2 Finite Dimensional Real Euclidean Jordan Algebras

In this section we present the concepts on Euclidean Jordan algebras which are relevant for our work. A more detailed exposition can be found in the monograph by Faraut and Korányi [5], and in Koecher's lecture notes, [10].

Let $\mathscr{A}$ be a finite dimensional real algebra with a bilinear mapping $(u, v) \mapsto u \cdot v$ from $\mathscr{A} \times \mathscr{A}$ into $\mathscr{A}$. Then $\mathscr{A}$ is a real Jordan algebra if $u \cdot v = v \cdot u$ and $u \cdot (u^2 \cdot v) = u^2 \cdot (u \cdot v)$, where $u^2 = u \cdot u$. From now on we suppose that if $\mathscr{A}$ is a real Jordan algebra, then $\mathscr{A}$ is a finite dimensional real algebra and has a unit element denoted by $\mathbf{e}$.

Example 1 The real vector space of real symmetric matrices of order n, $\mathscr{A} = \mathrm{Sym}_n(\mathbb{R})$, equipped with the bilinear map $u \bullet v = (uv + vu)/2$ is a real Jordan algebra.

Remark 1 Let $\mathscr{A}$ be a finite dimensional associative real algebra with the bilinear map $(u, v) \mapsto u \cdot v$. We introduce on $\mathscr{A}$ a structure of Jordan algebra by considering a new product $\bullet$ defined by $u \bullet v = (u \cdot v + v \cdot u)/2$ for all u and v in $\mathscr{A}$. The product $\bullet$ is called the Jordan product.

A real Jordan algebra is not necessarily an associative algebra. But, a real Jordan algebra is always power associative, that is, is an algebra such that the algebra spanned by any element and the unit is associative.

Let $\mathscr{A}$ be a n-dimensional real Jordan algebra and u in $\mathscr{A}$. The rank of u is the least natural number k such that $\{\mathbf{e}, u, \ldots, u^k\}$ is linearly dependent and we write $\mathrm{rank}(u) = k$. Since $\mathrm{rank}(u) \leq n$ we define the rank of $\mathscr{A}$ as being the natural number $\mathrm{rank}(\mathscr{A}) = \max\{\mathrm{rank}(u) : u \in \mathscr{A}\}$. An element u in $\mathscr{A}$ is regular if $\mathrm{rank}(u) = \mathrm{rank}(\mathscr{A})$. Let u be a regular element of $\mathscr{A}$ and $r = \mathrm{rank}(u)$. Then, there exist real scalars $a_1(u), a_2(u), \ldots, a_{r-1}(u)$ and $a_r(u)$ such that

$$u^r - a_1(u)u^{r-1} + \cdots + (-1)^r a_r(u)\mathbf{e} = 0, \tag{1}$$

where 0 is the null vector of $\mathscr{A}$. Taking into account (1) we conclude that the polynomial

$$p_u(\lambda) = \lambda^r - a_1(u)\lambda^{r-1} + \cdots + (-1)^r a_r(u) \tag{2}$$

is the minimal polynomial of u. When u is not regular the minimal polynomial of u has a degree less than r. The roots of the minimal polynomial of u are the eigenvalues of u.

A real Euclidean Jordan algebra $\mathscr{A}$ is a Jordan algebra with an inner product $< \cdot, \cdot >$ such that $< u \cdot v, w >=< v, u \cdot w >$ for all u, v and w in $\mathscr{A}$.

Example 2 The real vector space $\mathrm{Sym}_n(\mathbb{R})$ is a real Euclidean Jordan algebra when endowed with the Jordan product and with the inner product $< u, v >= \mathrm{tr}(uv)$, where tr denotes the usual trace of matrices.

Let $\mathscr{A}$ be a real Euclidean Jordan algebra with unit element $\mathbf{e}$. An element f in $\mathscr{A}$ is an idempotent if $f^2 = f$. Two idempotents f_1 and f_2 are orthogonal if $f_1 \cdot f_2 = 0$. A *complete system of orthogonal idempotents* of $\mathscr{A}$ is a set $\{f_1, f_2, \ldots, f_k\}$ such that (i) $f_i^2 = f_i, \forall i \in \{1, \ldots, k\}$; (ii) $f_i \circ f_j = 0, \forall i \neq j$ and (iii) $f_1 + f_2 + \cdots + f_k = \mathbf{e}$. An idempotent f is primitive if it is a nonzero idempotent of $\mathscr{A}$ and if it can't be written as a sum of two non-zero orthogonal idempotents. We say that $\{f_1, f_2, \ldots, f_k\}$ is a Jordan frame if $\{f_1, f_2, \ldots, f_k\}$ is a complete system of orthogonal idempotents such that each idempotent is primitive.

Theorem 1 *([5], p. 43) Let $\mathscr{A}$ be a real Euclidean Jordan algebra. Then for u in $\mathscr{A}$ there exists unique real numbers $\lambda_1, \lambda_2, \ldots, \lambda_k$, all distinct, and a unique complete system of orthogonal idempotents $\{f_1, f_2, \ldots, f_k\}$ such that*

$$u = \lambda_1 f_1 + \lambda_2 f_2 + \cdots + \lambda_k f_k. \tag{3}$$

The numbers λ_j's of (3) are the eigenvalues of u and the decomposition (3) is the first spectral decomposition of u.

Theorem 2 *([5], p. 44) Let $\mathscr{A}$ be a real Euclidean Jordan algebra with rank $(\mathscr{A}) = r$. Then, for each u in $\mathscr{A}$ there exists a Jordan frame $\{f_1, f_2, \ldots, f_r\}$ and real numbers $\lambda_1, \ldots, \lambda_{r-1}$ and λ_r such that*

$$u = \lambda_1 f_1 + \lambda_2 f_2 + \cdots + \lambda_r f_r. \tag{4}$$

The numbers λ_j's (with their multiplicities) are uniquely determined by u.

The decomposition (4) is called the second spectral decomposition of u. Regard that the second spectral decomposition of u is not unique.

3 Preliminaries on Strongly Regular Graphs

Herein we will introduce some relevant preliminaries on the theory of strongly regular graphs. Detailed information can be found in [8].

Along this paper, we consider only non-empty, not complete, undirected, simple graphs. Considering a graph G, we denote its vertex set by $V(G)$ and its edge set by $E(G)$. An edge of G with endpoints x and y is denoted by xy. In this case the vertices are called adjacent or neighbors. The number of vertices of G, $|V(G)|$, is called the order of G. If all vertices of G have k neighbors, then G is a k-regular graph.

Let G be a graph of order n. Then G is an (n, k, a, c)-strongly regular graph if it is k-regular and any pair of adjacent vertices have a common neighbors and any pair of non-adjacent vertices have c common neighbors. The parameters of an (n, k, a, c)-strongly regular graph are not independent and are related by the equality

$$k(k - a - 1) = (n - k - 1)c. \tag{5}$$

The adjacency matrix of G, $A = \left[a_{ij}\right]$, is a matrix of order n such that $a_{ij} = 1$, if the vertex i is adjacent to j and 0 otherwise. The adjacency matrix of a strongly regular graph satisfies the equation $A^2 = kI_n + aA + c(J_n - A - I_n)$, where J_n is the all one matrix of order n and I_n is the identity matrix of order n.

Equation (5) is an example of a condition that must be satisfied by the parameters of any strongly regular graph. Among the most important feasibility conditions there are the Krein conditions obtained in 1973 by Scott Jr [15], and the Absolute Bounds by Seidel, [4]. However, there are still many parameter sets for which we do not know if they correspond to a strongly regular graph. The most notable example is perhaps the four graph of Moore with parameter set (3250, 57, 0, 1). In this work we deduce some new inequalities on the parameters and on the spectra of a strongly regular graph.

4 Alternating Hadamard Series and Some Theorems on Strongly Regular Graphs

Let G be an (n, k, a, c)-strongly regular graph and A be its adjacency matrix with three distinct eigenvalues, namely k, θ and τ, and let $\mathscr{A} = \mathrm{Sym}_n(\mathbb{R})$. We consider the Euclidean Jordan subalgebra of $\mathscr{A}$, $\mathscr{A}^*$, see [8, p. 177], spanned by I_n, and the natural powers of A. Since A has three distinct eigenvalues, then $\mathscr{A}^*$ is a three dimensional real Euclidean Jordan algebra with $\mathrm{rank}(\mathscr{A}^*) = 3$. Let $\mathscr{B} = \{E_1, E_2, E_3\}$ be the unique complete system of orthogonal idempotents of $\mathscr{A}^*$ associated to A, with

$$\begin{aligned}
E_1 &= \frac{1}{n}I_n + \frac{1}{n}A + \frac{1}{n}(J_n - A - I_n),\\
E_2 &= \frac{|\tau|n + \tau - k}{n(\theta - \tau)}I_n + \frac{n + \tau - k}{n(\theta - \tau)}A + \frac{\tau - k}{n(\theta - \tau)}(J_n - A - I_n),\\
E_3 &= \frac{\theta n + k - \theta}{n(\theta - \tau)}I_n + \frac{-n + k - \theta}{n(\theta - \tau)}A + \frac{k - \theta}{n(\theta - \tau)}(J_n - A - I_n).
\end{aligned}$$

Let $M_n(\mathbb{R})$ be the set of square matrices of order n with real entries. For $B = [b_{ij}]$, $C = [c_{ij}]$ in $M_n(\mathbb{R})$, we denote by $B \circ C = [b_{ij}c_{ij}]$ the Hadamard product of matrices B and C (see [9]).

For B in $M_n(\mathbb{R})$ and for l in $\mathbb{N}$ we denote by $B^{\circ l}$ the Hadamard power of order l of B, respectively, with $B^{\circ 1} = B$.

Consider the idempotent E_3 given in the previous section. The eigenvalues q_1, q_2 and q_3 of E_3 are given by

$$\begin{aligned}
q_1 &= \frac{\theta n + k - \theta}{n(\theta - \tau)} + \frac{-n + k - \theta}{n(\theta - \tau)}k + \frac{k - \theta}{n(\theta - \tau)}(n - k - 1),\\
q_2 &= \frac{\theta n + k - \theta}{n(\theta - \tau)} + \frac{-n + k - \theta}{n(\theta - \tau)}\theta + \frac{k - \theta}{n(\theta - \tau)}(-\theta - 1),\\
q_3 &= \frac{\theta n + k - \theta}{n(\theta - \tau)} + \frac{-n + k - \theta}{n(\theta - \tau)}\tau + \frac{k - \theta}{n(\theta - \tau)}(-\tau - 1).
\end{aligned}$$

From E_3 we build the following partial sum:

$$S_{4l-1} = \sum_{j=1}^{2l}(-1)^{j-1}\frac{(E_3^{\circ 2})^{\circ(2j-1)}}{(2j-1)!} + \frac{1}{3!}\left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^3 \frac{1 - \left(\frac{\theta n+k-\theta}{n(\theta-\tau)}\right)^{4l}}{1 - \left(\frac{\theta n+k-\theta}{n(\theta-\tau)}\right)^{4}} I_n.$$

Since $\mathscr{A}^*$ is closed under the Hadamard product and $\mathscr{B}$ is a basis of $\mathscr{A}^*$, we can write S_{4l-1} as: $S_{4l-1} = \sum_{i=1}^{3} q^i_{S_{4l-1}} E_i$, where the $q^i_{S_{4l-1}}$ with $i \in \{1, 2, 3\}$, are the eigenvalues of S_{4l-1}. We prove that $q^i_{S_{4l-1}} \geq 0$, $\forall i \in \{1, 2, 3\}$. First, we note the following identity regarding one of the eigenvalues of $E_3^{\circ 2}$:

$$\left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^2 + \left(\frac{-n + k - \theta}{n(\theta - \tau)}\right)^2 k + \left(\frac{(k - \theta)}{n(\theta - \tau)}\right)^2 (n - k - 1) = \frac{\theta n + k - \theta}{n(\theta - \tau)}.$$

Secondly, since all of the eigenvalues of E_3 are smaller in modulus than q_1, then the eigenvalues of the summands of $\sum_{j=1}^{2l}(-1)^{j-1}(E_3^{\circ 2})^{\circ(2j-1)}/(2j-1)!$ when $j - 1$ is odd are smaller, in modulus, than $\frac{1}{3!}\left(\theta n + k - \theta)/(n(\theta - \tau))\right)^{2j-1}$. For this assertion we also use the property $\lambda_{\max}(A_1 \circ \cdots \circ A_i) \leq \lambda_{\max}(A_1) \ldots \lambda_{\max}(A_i)$, where $\lambda_{\max}(A)$ denotes the maximum eigenvalue of the matrix A. Therefore, we conclude that all the eigenvalues of S_{4l-1} are nonnegative. Now we consider the sum $S_\infty = \lim_{l\to\infty} S_{4l-1}$. Therefore we have:

$$S_\infty = \left[\sin\left(\frac{(\theta n + k - \theta)}{n(\theta - \tau)}\right)^2 + \frac{1}{3!}\left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^3 \frac{1}{1 - \left(\frac{\theta * n + k - \theta}{n(\theta - \tau)}\right)^4} \right] I_n +$$
$$+ \sin\left(\frac{-n + k - \theta}{n(\theta - \tau)}\right)^2 A + \sin\left(\frac{k - \theta}{n(\theta - \tau)}\right)^2 (J_n - A - I_n).$$

Let q^i_∞, $i \in \{1, 2, 3\}$ be the eigenvalues of S_∞ such that $S_\infty = \sum_{i=1}^3 q^i_\infty E_i$. Then, since $q^i_\infty = \lim_{l \to} q^i_{S_{2l-1}}$, for $i \in \{1, 2, 3\}$, and $q^i_{S_{2l-1}} \geq 0, \forall i \in \{1, 2, 3\}$, we conclude that $q^i_\infty \geq 0, \forall i \in \{1, 2, 3\}$.

Finally, we consider the new matrix, S_3, obtained as $S_3 = E_3 \circ S_\infty$. The eigenvalues of S_3 are also nonnegative because of the non-negativity of the eigenvalues of E_3 and S_∞ and the property $\lambda_{\min}(A \circ B) \geq \lambda_{\min}(A)\lambda_{\min}(B)$, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of the matrix A. From the non-negativity of the eigenvalues of S_3 we establish the following result.

Theorem 3 *Let X be a strongly regular graph with parameter set (n, k, a, c) and three distinct eigenvalues, k, θ and τ. If $k < n/3$ and $\theta < |\tau| - 2/3$, then*

$$k \leq \frac{56}{9} \frac{(3\theta + 1)^3 \theta^4}{32\theta^4 - 1}. \tag{6}$$

Proof Let q^i_3, $i \in \{1, 2, 3\}$ be the eigenvalues of S_3 then $S_3 = \sum_{i=1}^3 q^i_3 E_i$. We have already proved that all the eigenvalues of S_3 are nonnegative. In particular, we have that $q^1_3 \geq 0$, that is

$$0 \leq \frac{\theta n + k - \theta}{n(\theta - \tau)} \left[\sin\left(\frac{(\theta n + k - \theta)}{n(\theta - \tau)}\right)^2 + \frac{1}{3!}\left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^3 \frac{1}{1 - \left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^4} \right]$$
$$+ \frac{-n + k - \theta}{n(\theta - \tau)} \sin\left(\frac{-n + k - \theta}{n(\theta - \tau)}\right)^2 k +$$
$$+ \frac{k - \theta}{n(\theta - \tau)} \sin\left(\frac{k - \theta}{n(\theta - \tau)}\right)^2 (n - k - 1). \tag{7}$$

Since, for any strongly regular graph, we have $q_1 = 0$, then inequality (7) can be rewritten as

$$0 \leq \frac{\theta n + k - \theta}{n(\theta - \tau)} \left[\sin\left(\frac{(\theta n + k - \theta)}{n(\theta - \tau)}\right)^2 - \sin\left(\frac{k - \theta}{n(\theta - \tau)}\right)^2 \right]$$
$$+ \frac{1}{3!}\left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^4 \frac{1}{1 - \left(\frac{\theta n + k - \theta}{n(\theta - \tau)}\right)^4} +$$

$$+\frac{-n+k-\theta}{n(\theta-\tau)}\left[\sin\left(\frac{-n+k-\theta}{n(\theta-\tau)}\right)^2 k-\sin\left(\frac{k-\theta}{n(\theta-\tau)}\right)^2\right]k. \tag{8}$$

Applying the Mean Value Theorem, see [1, Theorem 4.8.2, p. 308], to (8) to the function sin in the interval $\left[((k-\theta)/(n(\theta-\tau)))^2,\ \ ((n-k+\theta)/(n(\theta-\tau)))^2\right]$ and after making the minorization of cos in this interval, and finally since $\sin\big((\theta n+k-\theta)/(n(\theta-\tau))\big)^2 \le \big((\theta n+k-\theta)/(n(\theta-\tau))\big)^2$ one obtains the equality (9).

$$0\le(\frac{\theta n+k-\theta}{n(\theta-\tau)})^3+\frac{1}{3!}\left(\frac{\theta n+k-\theta}{n(\theta-\tau)}\right)^4\frac{1}{1-\left(\frac{\theta n+k-\theta}{n(\theta-\tau)}\right)^4}+$$
$$+\frac{-n+k-\theta}{n(\theta-\tau)}\cos(\frac{(n-k+\theta)^2}{n(\theta-\tau)})^2\frac{1}{\theta-\tau}\frac{n-2k+2\theta}{n(\theta-\tau)}k. \tag{9}$$

Since $\theta<|\tau|-\frac{2}{3}$ implies that $\big((\theta n+k-\theta)/(n(\theta-\tau))\big)/\big(1-\big((\theta n+k-\theta)/(n(\theta-\tau))\big)^4\big)\le 1$ and finally since $\cos\big((n-k+\theta)/(n(\theta-\tau))\big)^2\ge 1-(1/32)(1/\theta^4)$ we obtain from (9) the inequality (10).

$$0\le\frac{7}{6}\left(\frac{\theta n+k-\theta}{n(\theta-\tau)}\right)^3+\frac{-n+k-\theta}{n(\theta-\tau)}\frac{32\theta^4-1}{32\theta^4}\frac{1}{\theta-\tau}\frac{n-2k+2\theta}{n(\theta-\tau)}k. \tag{10}$$

Using the fact that $k<n/3$ and making an algebraic manipulation on the right member of (10) we obtain $k\le 7\big(3\theta+1\big)^3\big(32\theta^4\big)36\big(32\theta^4-1\big)$. □

From Theorem 3 we obtain the Corollary 1.

Corollary 1 *Let X be an strongly regular with the distinct eigenvalues* θ, τ *and* k. *If* $k>\frac{2n}{3}-1$ *and* $|\tau|<\theta+\frac{4}{3}$ *then*

$$n-k-1\le\frac{56}{9}\frac{(3|\tau|-2)^3(|\tau|-1)^4}{32(|\tau|-1)^4-1}. \tag{11}$$

5 Numerical Results

In this section we present some examples of parameter sets that show the effectiveness of the deduced inequalities (6) and (11).

We present in Table 1 some examples of parameter sets (n,k,a,c) that do not verify the inequality (6) of Theorem 3. We consider the parameter sets $P_1=(64,21,0,3)$, $P_2=(300,92,10,36)$, $P_3=(1156,275,18,80)$, $P_4=(1225,408,59,174)$ and $P_5=(1225,352,24,132)$. For each example we have $k<n/3$ and we present the respective eigenvalues θ, τ and the value of $q_{\theta k}$ defined by

Table 1 Numerical results when $k < n/3$

	P_1	P_2	P_3	P_4	P_5
θ	1	2	3	2	2
τ	−11	−28	−65	−117	−110
$q_{\theta k}$	−8.2	−25.2	−80.5	−342	−286.2

Table 2 Numerical results when $k > 2n/3 - 1$

	P_6	P_7	P_8	P_9	P_{10}
θ	10.0	27.0	64.0	116.0	109.0
τ	−2.0	−3	−4	−3.0	−3
$q_{\tau k}$	−8.2	−25.2	−80.5	−342.2	−285.2

$$q_{\theta k} = \left[56(3\theta + 1)^3\theta^4\right] / \left[9(32\theta^4 - 1)\right] - k.$$

Next, in Table 2, we present some examples of parameter sets (n, k, a, c) that do not verify the inequality (11) of Corollary 1. We consider the parameter sets $P_6 = (64, 42, 30, 22)$, $P_7 = (300, 207, 150, 126)$, $P_8 = (1156, 880, 684, 624)$, $P_9 = (1225, 816, 581, 468)$ and $P_{10} = (1225, 872, 651, 545)$. For each example we have $k > 2n/3 - 1$ and we present the respective data as in Table 1 but in the last line we compute the value of $q_{\tau k} = \left[56(3|\tau| - 2)^3(|\tau| - 1)^4\right] / \left[9(32(|\tau| - 1)^4 - 1)\right] - (n - k - 1)$.

Acknowledgements Luís Vieira was supported by the European Regional Development Fund Through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEst- C/MAT/UI0144/2013.

References

1. Anton, H., Bivens, I., Davis, S.: Pacific Journal of Mathematics. Calculus, pp. 389–419. Wiley, New York (1963)
2. Cardoso, D.M., Vieira, L.: Euclidean Jordan algebras with strongly regular graphs. J. Math. Sci. **120**, 881–894 (2004)
3. Cardoso, D.M., Vieira, L.: On the optimal parameter of a self-concordant barrier over a symmetric cone. Eur. J. Oper. Res. **169**, 1148–1157 (2006)
4. Delsarte, Ph, Goethals, J.M., Seidel, J.J.: Bounds for system of lines and Jacobi polynomials. Philips Res. Rep. **30**, 91–105 (1975)
5. Faraut, J., Korányi, A.: Analysis on Symmetric Cones. Oxford Mathematical Monographs. Clarendon Press, Oxford (1994)
6. Faybusovich, L.: Linear systems in Jordan algebras and primal-dual interior-point algorithms. J. Comput. Appl. Math. **86**, 149–175 (1997)
7. Faybusovich, L.: Euclidean Jordan algebras and interior-point algorithms. Positivity. **1**, 331–357 (1997)
8. Godsil, C., Royle, G.: Algebraic Graph Theory. Chapman & Hall, New York (1993). On an algebraic generalization of the quantum mechanical formalism. Ann. Math. **35**, 29–64 (1934)
9. Horn, R., Jhonson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)

10. Koecher, M.: The Minnesota Notes on Jordan Algebras and Their Applications. Springer, Berlin (1999)
11. Mano, V.M., Vieira, L.A.: Admissibility conditions and asymptotic behavior of strongly regular graph. Int. J. Math. Model. Methods Appl. Sci. Methods **5**(6), 1027–1033 (2011)
12. Mano, V.M., Martins, E.A., Vieira, L.A.: A: feasibility conditions on the parameters of a strongly regular graph. Electron. Notes Discret. Math. **38**, 607–613 (2011)
13. Mano, V.M., Martins, E.A., Vieira, L.A.: A: generalized binomial series and strongly regular graphs. Proyecciones J. Math. **32**, 393–408 (2013)
14. Massam, H., Neher, E.: Estimation and testing for lattice condicional independence models on Euclidean Jordan algebras. Ann. Stat. **26**, 1051–1082 (1998)
15. Scott, Jr. L.L.: A condition on Higman parameters. Not. Am. Math. Soc. **20** A-97 (1973)
16. Vieira, L.A.: Euclidean Jordan algebras and inequalities on the parameters of a strongly regular graph. AIP Conf. Proc. **1168**, 995–998 (2009)

Actuarial Present Value and Variance for Changing Mortality and Stochastic Interest Rates

Bükre Yıldırım, A. Sevtap Selcuk-Kestel and N. Gülden Coşkun-Ergökmen

Abstract Stochastic modeling of interest rates is expected to lead a better risk management in long-term investments due to the rapid changes and random fluctuations in the economies. Considering the fact that deterministic interest rate approach does not yield realistic future values, a country-specific stochastic model is aimed to fit the interest rates based on the United States Treasury Inflation Protected Securities (TIPS) at 10-year constant maturity by using time series techniques. Under the assumption that interest rate follows an $ARMA(1, 1)$ model, the actuarial present value and its variance for a ten-year term life insurance policy are derived. Additionally, the stochastic mortality using Lee-Carter model for future mortality predictions is implemented to the U.S. Mortality tables over a period of 81 years. Based on these two stochastic patterns, the actuarial present value and the variance functions are calculated numerically for the years 2014 and forecasted for 2030. The accuracy of the proposed model is performed by assessing a comparative analysis with respect to a prespecified deterministic interest rate and mortality table.

Keywords Term-life insurance · Stochastic interest rate · Treasury securities · $ARMA(p, q)$ · Lee-Carter model · Actuarial present value · Actuarial variance

B. Yıldırım · A.S. Selcuk-Kestel (✉)
Middle East Technical University, Institute of Applied Mathematics,
Ankara, Turkey
e-mail: skestel@metu.edu.tr

B. Yıldırım
e-mail: bukre@metu.edu.tr

N.G. Coşkun-Ergökmen
Republic of Turkey, Prime Ministry, Undersecretariat Treasury,
Ankara, Turkey
e-mail: gulden.ergokmen@hazine.gov.tr

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_24

1 Introduction

The long-term valuation of policyholders and insurers accumulation requires a precise financial planning in insurance sector. The actuarial analyses are commonly based on the deterministic assumptions on interest and mortality rates. However, these important factors in evaluating major indicators such as net single premium, reserves, technical gains and annuities, play an important role in the actuarial valuation. A realistic approach is to include the impact of stochasticity in formulating interest rate which enables the actuary to express the path to be followed for the future time value of the money.

Many studies are available in the literature on stochastic interest and mortality rate modeling. Many of those also employ time series models to understand the behavior of the rates. Boyle [1] assumes that the force of interest is generated by a white noise series and autoregressive models of order one are introduced to model interest rates. Panjer and Bellhouse [2, 3] developed a general theory for both continuous and discrete models by using $AR(1)$ and $AR(2)$ processes to compute moments of insurance and annuity functions. Giacotto [4] analyzed present value functions with stochastic interest rates when the spot rates are modeled discrete or continuous stochastic processes. He modeled the interest rates when actuarial functions are considered by using stationary and nonstationary $ARIMA(p, 0, q)$ and $ARIMA(p, 1, q)$ processes. Dhaene [5] developed the study of Giacotto (1986); the force of interest is modelled as an $ARIMA(p, d, q)$ process. He used this model to compute the moments of present value functions. Frees [6, 7] examined the net premiums in life contingencies and extended the theory of life contingencies to a stochastic environment by using $MA(1)$ model. Parker [8] presented a model combining random interest rate and random future lifetimes for portfolios of identical life insurances by using Ornstein-Uhlenbeck process. Lai and Frees [9] studied the potential short-term consequences of changes in the interest rate environment by using linear and nonlinear $ARCH$ process. Zacks [10] investigated the accumulated value of some annuities-certain over a period of years with random interest rate. In Debicka's [11] work, the cash value of discrete-time payment streams in insurance contracts are calculated where the interest rate and future-lifetime are random.

For the continuous modeling of interest rates Merton [12] (1973) used Ito process for the first time. Later Vasicek [13] employed Ornstein-Uhlenbeck type of short rate model with mean-reverting characteristics of the data. Cox, Ingersoll, and Ross (CIR) [14] (1985) introduced their original model to eliminate the shortcoming of previous models which was mainly based on the positive probability of negative interest rates.

Also another approach for modeling interest rate movements are the use of autoregressive conditional heteroskedasticity (ARCH) models, introduced by Engle [15]. These models are developed into generalized autoregressive conditional heteroskedasticity (GARCH) by Bollerslev [16], and to exponential GARCH (EGARCH) by Nelson [17]. The significant point for these models are that they indicated volatility persistence in high degrees, which was the shortcoming in the

CIR model. These approaches are the most commonly interest rate modeling which takes into account the random volatility in the market.

This study aims to evaluate and derive the actuarial present value and its variance for a term life insurance under the assumption that interest rates follow $ARMA(1, 1)$. The motivation is to observe the influence of stochastic interest and mortaliy rates on the net single premium and its variance. A developed market in life insurance is taken into account to illustrate the impact and efficiency of the proposed approach. To achieve the proposed approach, monthly United States TIPS at 10-year constant maturity are taken into account and an appropriate model is fitted. The implementation of the model is done on a mortality table whose stochastic pattern is predicted using Lee-Carter model. Sensitivity checks are done by comparing deterministic interest and mortality rates with stochastic ones. All computations are performed using Matlab R2013a, and Microsoft Excel 2010. The derivations of the actuarial present value and the variance modified based on [18] are presented with their proofs.

The organization of the chapter is as follows: The basic model proposed is presented in the next section. Market yields on United States TIPS at 10-year constant maturity are analyzed and modeled through $ARMA(1, 1)$ model and actuarial present values and the variances are derived. The U.S. mortality rates between years 1933 and 2013 are used to estimate the future mortality rates for a period of next 16 years. A comparison of deterministic and stochastic approach on the present values is presented in the last section.

2 Stochastic Interest Rate Model

Treasury bills are the safest and highly demanded instruments in financial markets. Most of the life insurance regulations on the valuation of life insurance reserves require insurance companies to invest a portion of their accumulations on safe investments like treasury bills and government bonds. For this reason, to capture the behaviour of a risk free asset in a volatile market gains importance. As the first step in actuarial valuation under stochastic interest rate, we start defining a time dependent model using the U.S. TIPS at 10-year maturity. The data set is collected from the web page of Federal Reserve [19]. The nominal interest rates can be converted to their real equivalences as follows:

$$\begin{aligned} i_t =& (1 + monthly\ interest\ rate)^{1/m} - 1 \\ i'_t =& \frac{i_t - e_t}{1 + e_t}. \end{aligned} \tag{1}$$

Here, i_t, e_t and i'_t denote the monthly interest rate, the inflation rate and the real interest rate for the tth term, respectively. The monthly rates of a 12-year period starting from January 2003 to December 2015 are inflation adjusted. Therefore, a

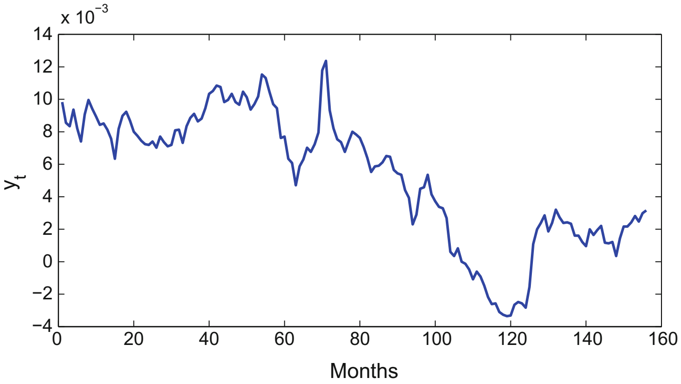


Fig. 1 The real interest rates of the U.S. TIPS at 10-year maturity between 2003 and 2015 [19]

log-transformation of the real interest rates rates are employed to fit an appropriate model.

The interest rates, y_t, plotted with respect to time in Fig. 1, illustrate the sharp decrease in years 2008 and 2013 and a declining trend over years. The preliminary descriptive analysis and frequency plot of returns yield a monthly average value of 0.0053; a median value of 0.0064; and a mode value of 0.0024. The standard deviation of the process is found to be 0.0040. The pattern of temporal dependence is analyzed through autocorrelated (ACF) and partial autocorrelated ($PACF$) functions which are presented in Fig. 2. It can be seen that the original series has a trend which require testing the existence of the unit root. ADF test statistics of the first differenced data (p value<0.01) assure the stationarity to proceed in model estimation.

The time series model, $ARIMA(1, 1)$ [20] is defined as

$$y_t = \delta + \varphi(y_{t-1} - \delta) + \varepsilon_t + \beta\varepsilon_{t-1} \tag{2}$$

where δ, φ and β denote the parameters corresponding to the drift, AR and MA coefficients, respectively. The model estimation yields the values for the parameters which are illustrated in Table 1. The p-values of the test statistics show that the coefficients are significant validating the model at the main step of the time series modeling, except the constant term. Besides the statistical justification of the parameters, the best fit in time series require the diagnostic tests on the residuals. The residual analysis yields a mean value which is almost equal to zero, -5.6887×10^{-6}, and a standard deviation of 7.9154×10^{-4}. As it can be seen in the Fig. 3, the ACF and $PACF$ graphs of residuals support that they are white noise and Q–Q plot justifies that the normality in residuals is justified.

The stationarity in ARMA models presented in Eq. (2) leads us to represent the linear model in terms of its residuals as given below:

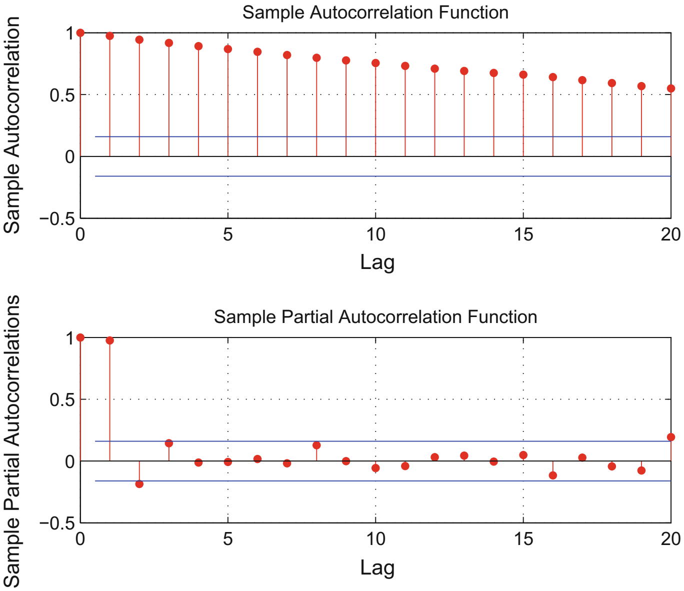


Fig. 2 ACF and $PACF$ plots of the log-returns

Table 1 Parameter estimates of the differenced data

Parameter	Value	Standard error	t statistic
Constant	0.000129	0.0001470	0.87733
AR(1)	0.969197	0.0201669	48.0587
MA(1)	0.24011	0.0879572	2.72985
Variance	6.22103e-07	2.77698e-07	2.24022

$$y_t = \delta + \varepsilon_t + (\varphi_1 + \beta_1) \sum_{j=1}^{\infty} \varphi_1^{j-1} \varepsilon_{t-j} \tag{3}$$

As it can be seen from Eq. (3), y_t could easily be estimated in terms of its residual terms (random error). Using this stationary linear model is convenient in the sense of deriving distribution of the series with respect to the moment generating functions, as the residuals follow normal distribution. Therefore, the moment generating function will be the basic term to derive the actuarial present value and its variance.

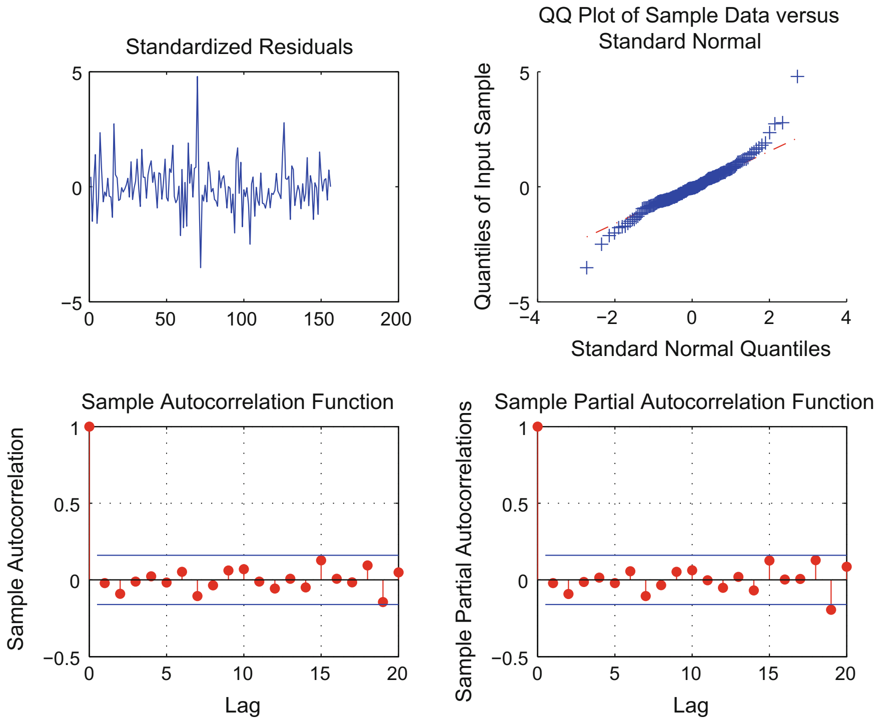


Fig. 3 Residual tests and normality check

3 Actuarial Present Value and Its Variance Under ARMA (1,1)

Actuarial present value for a whole life insurance which pays a pre-determined benefit at the end of the year of death is a function of the interest and the mortality rates. Given V_t denotes the interest discount factor from the time of payment back to the time of policy issued, the present value, Z, of the amount of the payment, b_t, is

$$Z = b_t V^{t+1}, \tag{4}$$

whose probability distribution can be expressed as

$$Pr(Z = b_t v^{t+1}) = \ {}_t p_x \ q_{x+t}. \tag{5}$$

Here, for a person aged x and having maximum lifetime till age w, ${}_t p_x$ represents the probability of living between ages x and $x + t$ and q_{x+t} shows the probability of dying between ages $x + t$ and $x + t + 1$.

Let A_x denote the actuarial present value for a whole life insurance issued on a person whose age is (x). For a life insurance policy, mortality cost for each year is computed separately and its aggregate constitutes the net single premium. It is expressed as

$$A_x = \sum_{0}^{w} V^{t+1}\ b_t\ {}_t p_x\ q_{x+t}\ dt \tag{6}$$

Another life insurance type, n-year Term Life, provides the benefit only if the insured dies within the n-years of issue date. We denote the actuarial present value for the n-year term insurance with a benefit b_t, as $A_{\bar{x}:\overline{n}|}$

$$A_{\bar{x}:\overline{n}|} = \sum_{0}^{n} V^{t+1}\ b_t\ {}_t p_x\ q_{x+t}\ dt. \tag{7}$$

As the discount function, V_t, is a function of continuous interest rate, the expected value, $E[V_t]$, and the variance, $\text{Var}[V_t]$ under the assumption of stochastic interest rate and its impact on net premium has to be determined.

Proposition 1 *Let y_t follow $ARMA(1,1)$ at time t given in Eq. (3). The present value of a single payment, V_n, for $t = n$ is*

$$\begin{aligned} V_n = \exp&\left(-n\delta - (\varphi_1+\beta_1)\varepsilon_0\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right) - \varepsilon_n\right) \times \\ &\exp\left(-\sum_{j=1}^{n-1}\varepsilon_j\left[1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right]\right) \times \\ &\exp\left(-(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\sum_{j=1}^{\infty}\varphi_1^j\varepsilon_{-j}\right). \end{aligned} \tag{8}$$

The parameters in V_n derived above are defined in Eq. (3).

Proof

$$V_n = \prod_{t=1}^{n}(1+i_t)^{-1} = \exp\left(-\sum_{t=1}^{n} y_t\right)$$

$$V_n = \exp\left[-\sum_{t=1}^{n}\left(\delta + \varepsilon_t + (\varphi_1+\beta_1)\sum_{j=1}^{\infty}\varphi_1^{j-1}\varepsilon_{t-j}\right)\right]$$

$$V_n = \exp\left[-\sum_{t=1}^{n}\delta - \sum_{t=1}^{n}\varepsilon_t - (\varphi_1+\beta_1)\sum_{t=1}^{n}\sum_{j=1}^{\infty}\varphi_1^{j-1}\varepsilon_{t-j}\right]$$

$$V_n = \exp\Big[- n\delta - (\varepsilon_1 + ... + \varepsilon_n) \\ - (\varphi_1 + \beta_1) \sum_{t=1}^{n} (\varepsilon_{t-1} + \varphi_1 \varepsilon_{t-2} + .. + \varphi_1^n \varepsilon_{t-n} + ...)\Big]$$

$$V_n = \exp\Big[- n\delta - (\varepsilon_1 + ... + \varepsilon_n) \\ - (\varphi_1 + \beta_1) \big[(\varepsilon_0 + ... + \varepsilon_{n-1}) + \varphi_1(\varepsilon_{-1} + ... + \varepsilon_{n-2}) + ...\big]\Big]$$

$$V_n = \exp\Big[- n\delta - \varepsilon_0(\varphi_1 + \beta_1)(1 + \varphi_1 + \varphi_1^2 + ... + \varphi_1^{n-1}) - \varepsilon_n \\ - \varepsilon_1(1 + (\varphi_1 + \beta_1)(1 + \varphi_1 + \varphi_1^2 + ... + \varphi_1^{n-2})) \\ - ... - \varepsilon_{-1}(\varphi_1 + \beta_1)(\varphi_1 + \varphi_1^2 + ... + \varphi_1^n) - ...\Big]$$

$$V_n = \exp\Big[- n\delta - \varepsilon_0(\varphi_1 + \beta_1) \sum_{t=0}^{n-1} \varphi_1^t - \varepsilon_n - \varepsilon_1 \Big(1 + (\varphi_1 + \beta_1) \sum_{t=0}^{n-2} \varphi_1^t\Big) \\ - ... - \varepsilon_{-1}(\varphi_1 + \beta_1)\varphi_1 \sum_{t=0}^{n-1} \varphi_1^t - ...\Big]$$

$$V_n = \exp\Big[- n\delta - (\varphi_1 + \beta_1)\varepsilon_0 \left(\frac{1 - \varphi_1^n}{1 - \varphi_1}\right) - \varepsilon_n - \\ \sum_{j=1}^{n-1} \varepsilon_j \left[1 + (\varphi_1 + \beta_1) \sum_{t=0}^{n-j-1} \varphi_1^t\right] - (\varphi_1 + \beta_1) \left(\frac{1 - \varphi_1^n}{1 - \varphi_1}\right) \sum_{j=1}^{\infty} \varphi_1^j \varepsilon_{-j}\Big].$$

Proposition 2 *Given V_n defined in Eq. (8), the expected value becomes*

$$E(V_n) = e^{-n\delta} M(-1) \prod_{j=1}^{n-1} M \left[-\left(1 + (\varphi_1 + \beta_1) \prod_{t=0}^{n-j-1} \varphi_1^t\right)\right], \tag{9}$$

where M defines the moment generating function.

Proof

$$E(V_n) = E\Big[\exp\Big(- n\delta - (\varphi_1 + \beta_1)\varepsilon_0 \left(\frac{1 - \varphi_1^n}{1 - \varphi_1}\right) - \varepsilon_n$$

$$-\sum_{j=1}^{n-1}\varepsilon_j\left[1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right]$$
$$-(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\sum_{j=1}^{\infty}\varphi_1^j\varepsilon_{-j}\Bigg)\Bigg]$$

$$E(V_n)=e^{-n\delta}e^{-\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)(\varphi_1+\beta_1)\varepsilon_0}\times M(-1)\times$$
$$E\left[\exp\left(-\sum_{j=1}^{n-1}\varepsilon_j\left[1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right]\right)\right]\times$$
$$E\left[\exp\left(-(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\sum_{t=1}^{\infty}\varphi_1^t\varepsilon_{-t}\right)\right]$$
$$E(V_n)=e^{-n\delta}e^{-(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\varepsilon_0}\times M(-1)\times$$
$$\prod_{j=1}^{n-1}M\left[-\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right)\right]\times$$
$$\prod_{t=1}^{\infty}M\left[-(\varphi_1+\beta)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\varphi_1^t\right].$$

Here, $\varepsilon_t \sim N(0,\sigma_\varepsilon^2)$ and $M(t)=\exp(\sigma_\varepsilon^2 t^2/2)$. Considering $\varepsilon_0=0$ and $-1<\varphi_1<1$, the last term in proof is taken as equal to 1 [21]. Using this property,

$$\prod_{t=1}^{\infty}M\left[-(\varphi_1+\beta)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\varphi_1^t\right]=1 \tag{10}$$

the expected value of the present value finally is derived as

$$E(V_n)=e^{-n\delta}M(-1)\prod_{j=1}^{n-1}M\left[-\left(1+(\varphi_1+\beta_1)\prod_{t=0}^{n-j-1}\varphi_1^t\right)\right]. \tag{11}$$

Proposition 3 *The actuarial present value of n-year term-life insurance, $A_{\bar{x}:\overline{n}|}$, under ARMA(1,1) stochastic interest rate assumption is derived as*

$$
\begin{aligned}
A_{\bar{x}:\overline{n}|} &= E[V_{K+1}] = E_V E_{K|V}[V_{K+1}] \\
A_{\bar{x}:\overline{n}|} &= E_V[\sum_{k=0}^{n} V_{k+1\,k}p_x.q_{x+k}] \\
A_{\bar{x}:\overline{n}|} &= C_1 \sum_{k=0}^{n} e^{-(k+1)\delta}{}_kp_x.q_{x+k}
\end{aligned}
\tag{12}
$$

where $C_1 = e^{(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\varepsilon_0} M(-1)\prod_{j=1}^{n-1} M\left[-\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right)\right]$.
Taking $\varepsilon_0 = 0$ *and letting* $n = k+1$, C_1 *becomes*

$$
\begin{aligned}
C_1 &= M(-1)\prod_{j=1}^{n-1} M\left[-\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right)\right] \\
C_1 &= M(-1)\prod_{j=1}^{k} M\left[-\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{k-j}\varphi_1^t\right)\right].
\end{aligned}
\tag{13}
$$

Proposition 4 *The variance of* $A_{\bar{x}:\overline{n}|}$ *is derived as*

$$
Var(A_{\bar{x}:\overline{n}|}) = C_2 \sum_{k=0}^{n} e^{-2(k+1)\delta}{}_kp_x.q_{x+k} - (A_{\bar{x}:\overline{n}|})^2 \tag{14}
$$

Here, $C_2 = e^{-2(\varphi_1+\beta_1)\left(\frac{1-\varphi_1^n}{1-\varphi_1}\right)\varepsilon_0} M(-2)\prod_{j=1}^{n-1} M\left[-2\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{n-j-1}\varphi_1^t\right)\right]$.
Taking $\varepsilon_0 = 0$ *and* $n = k+1$, *is simplified to*

$$
C_2 = M(-2)\prod_{j=1}^{k} M\left[-2\left(1+(\varphi_1+\beta_1)\sum_{t=0}^{k-j}\varphi_1^t\right)\right] \tag{15}
$$

The propositions given above yields a formulation on the actuarial valuation of n-year term insurance under stochastic interest rate.

4 The Impact of Stochastic Mortality on $A_{\bar{x}:\overline{n}|}$

A life table shows fundamental parameters of a population for each age or age group, such as; the number of survivors, the number of deaths, the probability that they die or live to their next birthday and the life expectancy. It describes the mortality and survival pattern of a population. Mortality data and life tables, originate from observations concerning a whole national population or a specific part of a population (e.g. retired workers, disabled people, etc.) or an insurers portfolio, and so on. The past life table data does not assure its future outcome. Hence in order to price insurance products properly, actuaries must use projections of future insured events. To do this, actuaries developed mathematical models for estimating the mortality. The assumptions on constructing life tables require the observation of a closed group which needs many years to come up with an accurate estimate of the death rates. For this reason, the yearly population census, death and birth rates are employed to construct an efficient model. Additionally, the improvement in technology, medicine result in increase in the expected lifetime. Therefore, the time change on the mortalities show also a stochastic pattern. One of the models which takes into account the time influence and allows a good prediction on the future mortalities is Lee-Carter Model (LC) [22]. It expresses the mortality as a probability process and it is one of the most commonly used one in the literature to forecast the future mortality. It is essentially built for life expectancy forecasting, but can also be used for mortality forecasting.

LC model defines the force of mortality, μ_{x+t}, at age x and at time t using parameters α_x, β_x, and k_t as

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t} \qquad \varepsilon_{x,t} \sim N(0, \sigma) \tag{16}$$

where $m_{x,t}$ represents the central death rate, α_x is the average level of mortality at each age, β_x shows the sensitivity to κ_t at different ages, κ_t shows the general speed of mortality improvement over time, and $\varepsilon_{x,t}$ explains the error term and captures the remaining variations under the conditions

$$\begin{aligned} \sum \beta_x &= 1 \\ \sum \kappa_t &= 0. \end{aligned} \tag{17}$$

The U.S. Mortality rates between years 1933 and 2014 retrieved from open source internet website [23] are utilized for estimating the following 16-year mortalities using the LC model. The time behavior of the parameter and mortality estimates are presented in Fig. 4. The parameters on estimating the central death rates of US total population for ages 0–110, over the chosen period are illustrated in top two and bottom-right figures. Based on these estimates, the behavior of mortality rates is shown in the lower-right graph. The parameter α_x has an increasing rate by age, whereas, β_x and κ_t both decrease with respect to age. The time influence on the change on mortality rates can be observed in Fig. 5. The difference between the survival probability in 2014 and the forecast in 2030 present is observable, especially, between ages 80–100.

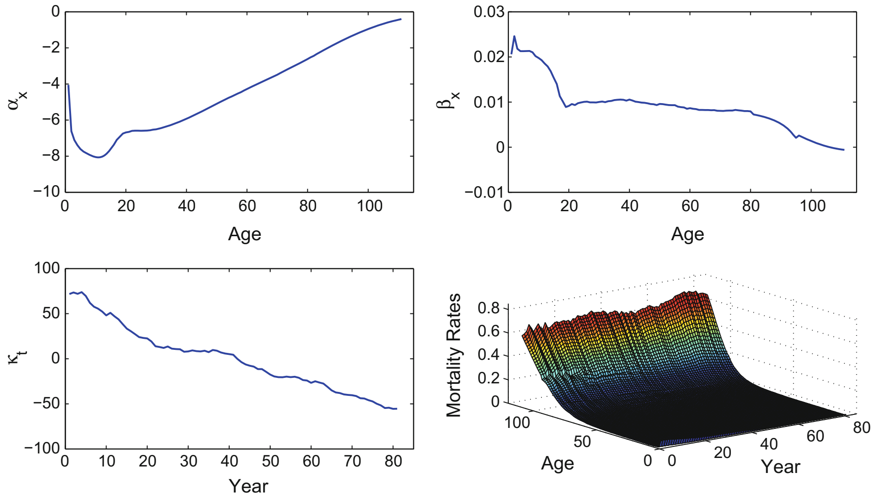


Fig. 4 Parameter estimates for the U.S mortality tables using Lee-Carter model

Fig. 5 Survival probability and its prediction using *LC* model for 2014 and 2030, repectively

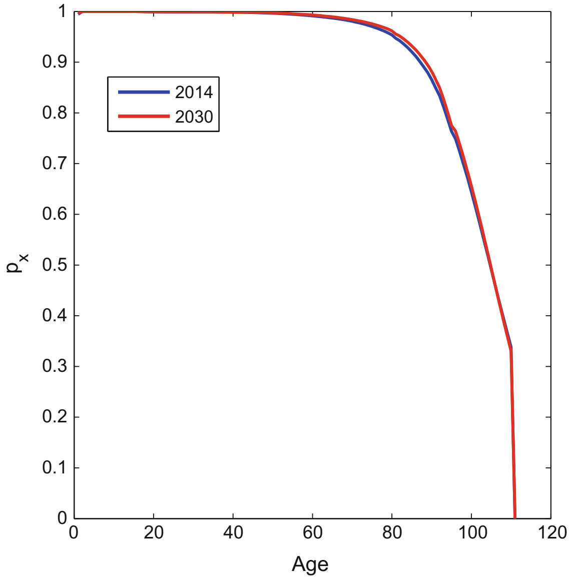


Table 2 Actuarial present value and its variance for deterministic (D) and stochastic (S) approaches

2014					2030				
x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}	x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}
0	0.0058	0.0051	0.0067	0.0066	0	0.0036	0.0032	0.0043	0.0043
1	0.0010	0.0007	0.0014	0.0014	1	0.0007	0.0005	0.0011	0.0011
2	0.0008	0.0006	0.0012	0.0012	2	0.0006	0.0004	0.0010	0.0010
3	0.0007	0.0005	0.0011	0.0011	3	0.0006	0.0004	0.0010	0.0010
4	0.0007	0.0004	0.0011	0.0011	4	0.0006	0.0004	0.0010	0.0010
5	0.0007	0.0005	0.0012	0.0012	5	0.0006	0.0004	0.0010	0.0010
6	0.0008	0.0005	0.0014	0.0014	6	0.0007	0.0004	0.0011	0.0011
7	0.0009	0.0006	0.0017	0.0017	7	0.0008	0.0005	0.0013	0.0013
8	0.0012	0.0007	0.0022	0.0022	8	0.0009	0.0005	0.0016	0.0016
9	0.0015	0.0008	0.0028	0.0028	9	0.0011	0.0006	0.0021	0.0021
10	0.0018	0.0010	0.0034	0.0034	10	0.0014	0.0008	0.0026	0.0026
11	0.0022	0.0012	0.0041	0.0041	11	0.0016	0.0009	0.0031	0.0031
12	0.0026	0.0015	0.0048	0.0048	12	0.0019	0.0011	0.0036	0.0036
13	0.0031	0.0018	0.0055	0.0055	13	0.0023	0.0013	0.0041	0.0041
14	0.0035	0.0021	0.0061	0.0060	14	0.0026	0.0016	0.0046	0.0046
15	0.0040	0.0025	0.0067	0.0066	15	0.0030	0.0019	0.0051	0.0051
16	0.0044	0.0028	0.0072	0.0071	16	0.0033	0.0021	0.0055	0.0055
17	0.0047	0.0031	0.0076	0.0075	17	0.0036	0.0024	0.0058	0.0057
18	0.0049	0.0032	0.0078	0.0077	18	0.0038	0.0025	0.0060	0.0059
19	0.0049	0.0033	0.0079	0.0078	19	0.0038	0.0025	0.0061	0.0060
20	0.0050	0.0034	0.0080	0.0079	20	0.0039	0.0026	0.0062	0.0061
21	0.0051	0.0034	0.0081	0.0080	21	0.0039	0.0026	0.0063	0.0062
22	0.0051	0.0034	0.0082	0.0081	22	0.0040	0.0026	0.0064	0.0063
23	0.0052	0.0034	0.0084	0.0083	23	0.0041	0.0027	0.0066	0.0065
24	0.0053	0.0035	0.0086	0.0085	24	0.0042	0.0027	0.0068	0.0067
25	0.0054	0.0035	0.0088	0.0087	25	0.0043	0.0028	0.0070	0.0069
26	0.0056	0.0036	0.0091	0.0090	26	0.0045	0.0029	0.0073	0.0072
27	0.0058	0.0037	0.0095	0.0094	27	0.0046	0.0030	0.0076	0.0075
28	0.0060	0.0039	0.0099	0.0098	28	0.0049	0.0032	0.0080	0.0079
29	0.0064	0.0041	0.0105	0.0104	29	0.0051	0.0033	0.0084	0.0083
30	0.0068	0.0043	0.0112	0.0111	30	0.0053	0.0034	0.0088	0.0087
31	0.0072	0.0046	0.0119	0.0118	31	0.0056	0.0036	0.0093	0.0092
32	0.0077	0.0049	0.0128	0.0126	32	0.0060	0.0038	0.0099	0.0098
33	0.0082	0.0051	0.0137	0.0135	33	0.0064	0.0040	0.0106	0.0105
34	0.0088	0.0055	0.0148	0.0146	34	0.0069	0.0043	0.0115	0.0114
35	0.0096	0.0060	0.0161	0.0158	35	0.0074	0.0047	0.0125	0.0124

Table 2 (continued)

2014					2030				
x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}	x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}
36	0.0104	0.0065	0.0175	0.0171	36	0.0080	0.0050	0.0136	0.0134
37	0.0112	0.0070	0.0189	0.0186	37	0.0088	0.0055	0.0149	0.0147
38	0.0122	0.0076	0.0206	0.0202	38	0.0096	0.0059	0.0163	0.0160
39	0.0133	0.0082	0.0225	0.0219	39	0.0104	0.0065	0.0177	0.0174
40	0.0145	0.0089	0.0244	0.0238	40	0.0115	0.0071	0.0194	0.0190
41	0.0158	0.0097	0.0267	0.0259	41	0.0126	0.0078	0.0213	0.0208
42	0.0172	0.0106	0.0290	0.0281	42	0.0137	0.0085	0.0232	0.0227
43	0.0187	0.0115	0.0316	0.0306	43	0.0150	0.0093	0.0254	0.0247
44	0.0204	0.0125	0.0344	0.0332	44	0.0164	0.0101	0.0276	0.0269
45	0.0222	0.0136	0.0374	0.0360	45	0.0178	0.0110	0.0301	0.0292
46	0.0242	0.0148	0.0408	0.0391	46	0.0194	0.0119	0.0328	0.0317
47	0.0264	0.0160	0.0445	0.0425	47	0.0212	0.0130	0.0358	0.0345
48	0.0287	0.0174	0.0484	0.0460	48	0.0231	0.0141	0.0390	0.0374
49	0.0313	0.0188	0.0528	0.0500	49	0.0252	0.0153	0.0426	0.0408
50	0.0341	0.0205	0.0574	0.0541	50	0.0275	0.0167	0.0465	0.0443
51	0.0371	0.0221	0.0625	0.0586	51	0.0301	0.0181	0.0508	0.0482
52	0.0404	0.0240	0.0680	0.0633	52	0.0328	0.0197	0.0553	0.0523
53	0.0440	0.0260	0.0742	0.0686	53	0.0358	0.0213	0.0605	0.0568
54	0.0479	0.0281	0.0807	0.0741	54	0.0391	0.0232	0.0659	0.0616
55	0.0522	0.0304	0.0876	0.0799	55	0.0427	0.0252	0.0718	0.0666
56	0.0566	0.0328	0.0950	0.0860	56	0.0464	0.0273	0.0781	0.0720
57	0.0614	0.0354	0.1029	0.0922	57	0.0504	0.0295	0.0847	0.0775
58	0.0665	0.0381	0.1112	0.0988	58	0.0548	0.0319	0.0917	0.0833
59	0.0719	0.0408	0.1201	0.1057	59	0.0593	0.0343	0.0993	0.0894
60	0.0779	0.0438	0.1298	0.1129	60	0.0644	0.0370	0.1076	0.0960
61	0.0842	0.0469	0.1403	0.1206	61	0.0698	0.0397	0.1166	0.1029
62	0.0910	0.0502	0.1514	0.1284	62	0.0755	0.0426	0.1259	0.1100
63	0.0982	0.0534	0.1635	0.1367	63	0.0816	0.0455	0.1362	0.1176
64	0.1060	0.0568	0.1765	0.1453	64	0.0883	0.0487	0.1473	0.1255
65	0.1145	0.0604	0.1904	0.1541	65	0.0954	0.0520	0.1590	0.1337
66	0.1236	0.0641	0.2055	0.1632	66	0.1032	0.0554	0.1719	0.1423
67	0.1336	0.0681	0.2219	0.1726	67	0.1117	0.0591	0.1859	0.1513
68	0.1445	0.0722	0.2395	0.1821	68	0.1209	0.0630	0.2011	0.1606
69	0.1560	0.0762	0.2582	0.1915	69	0.1308	0.0669	0.2173	0.1700
70	0.1686	0.0805	0.2786	0.2009	70	0.1416	0.0711	0.2350	0.1797

Table 2 (continued)

2014					2030				
x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}	x	$A^{D}_{\bar{x}:\overline{10}\rceil}$	Var^{D}	$A^{S}_{\bar{x}:\overline{10}\rceil}$	Var^{S}
71	0.1826	0.0846	0.3020	0.2107	71	0.1539	0.0752	0.2559	0.1903
72	0.1980	0.0888	0.3270	0.2200	72	0.1676	0.0796	0.2788	0.2010
73	0.2144	0.0926	0.3541	0.2286	73	0.1824	0.0838	0.3036	0.2113
74	0.2324	0.0963	0.3835	0.2363	74	0.1987	0.0880	0.3308	0.2213
75	0.2521	0.0998	0.4152	0.2427	75	0.2169	0.0922	0.3607	0.2305
76	0.2732	0.1027	0.4489	0.2472	76	0.2366	0.0961	0.3929	0.2384
77	0.2960	0.1050	0.4848	0.2496	77	0.2583	0.0996	0.4277	0.2446
78	0.3205	0.1067	0.5225	0.2493	78	0.2818	0.1027	0.4649	0.2486
79	0.3466	0.1076	0.5616	0.2460	79	0.3072	0.1051	0.5040	0.2498
80	0.3742	0.1076	0.6021	0.2394	80	0.3346	0.1068	0.5455	0.2478
81	0.4021	0.1058	0.6431	0.2293	81	0.3626	0.1066	0.5881	0.2421
82	0.4312	0.1030	0.6841	0.2159	82	0.3922	0.1053	0.6319	0.2324
83	0.4611	0.0986	0.7257	0.1989	83	0.4234	0.1021	0.6778	0.2182
84	0.4916	0.0927	0.7669	0.1785	84	0.4561	0.0972	0.7247	0.1993
85	0.5224	0.0856	0.8067	0.1557	85	0.4896	0.0905	0.7710	0.1763
86	0.5517	0.0784	0.8411	0.1335	86	0.5214	0.0836	0.8106	0.1533
87	0.5800	0.0706	0.8722	0.1113	87	0.5526	0.0759	0.8469	0.1295
88	0.6074	0.0628	0.8996	0.0902	88	0.5829	0.0678	0.8793	0.1060
89	0.6338	0.0551	0.9231	0.0708	89	0.6124	0.0597	0.9074	0.0838
90	0.6586	0.0479	0.9427	0.0539	90	0.6403	0.0518	0.9310	0.0641
91	0.6815	0.0413	0.9584	0.0398	91	0.6662	0.0446	0.9501	0.0473
92	0.7033	0.0355	0.9707	0.0284	92	0.6906	0.0382	0.9651	0.0336
93	0.7226	0.0306	0.9799	0.0196	93	0.7122	0.0328	0.9763	0.0231
94	0.7394	0.0263	0.9866	0.0132	94	0.7305	0.0280	0.9843	0.0154
95	0.7534	0.0226	0.9913	0.0086	95	0.7451	0.0238	0.9900	0.0099
96	0.7680	0.0195	0.9946	0.0054	96	0.7611	0.0204	0.9939	0.0061
97	0.7814	0.0169	0.9967	0.0033	97	0.7757	0.0176	0.9964	0.0037
98	0.7937	0.0147	0.9981	0.0020	98	0.7891	0.0152	0.9979	0.0022
99	0.8050	0.0129	0.9989	0.0012	99	0.8014	0.0133	0.9988	0.0013
100	0.8154	0.0113	0.9993	0.0008	100	0.8126	0.0116	0.9993	0.0008

5 Sensitivity Analyses and Concluding Comments

The sensitivity of the actuarial present value and its variance to deterministic and stochastic approaches is performed according to two important variables which are taken into account in this study. The first one considers a deterministic annual interest rate of 9% (denoted D) and random interest rates following an $ARMA(1,1)$

(denoted S). The latter one compares the valuations with respect to the type of mortality table. Total (male and female) mortality table for the year 2014 and forecasted mortality table for 2030 using LC model are compared under stochastic and deterministic interest rate cases. Equations (12) and (14) are employed to quantify $A^{D}_{\bar{x}:\overline{10}|}$, $\mathrm{Var}(A^{D}_{\bar{x}:\overline{10}|})$, $A^{S}_{\bar{x}:\overline{10}|}$ and $\mathrm{Var}(A^{S}_{\bar{x}:\overline{10}|})$ which denote the expected values and variances for deterministic (D) and stochastic (S) cases, respectively. Table 2 summarizes these results with respect to the ages. Figures 6 and 7 are presented to expose the impact of stochastic approach on the valuation compared to the deterministic case. The variance of the actuarial present value is observed to be low for young and very old ages, however, high between ages 50 and 90 in both cases.

Based on the proposed approach, we conclude that the stochastic modeling of interest rates yields higher actuarial present values and variances for both deterministic and stocastic mortality approaches. Even though the maximum volatility is around 15% compared to the deterministic one, stochastic interest model results in more conservative approach in handling the risk which will result in higher premium rates. This may be a discouraging temptation in marketing life insurance products, however, it reduces the adverse selection.

As the life insurance products are long-term investments for both insurer and insured, the deterministic assumptions will not be realistic, especially considering the longevity risk and non-stationary financial markets. This study enables researchers and insurance experts to quantify the risk to be taken if the actuarial valuation is done under stochastic framework and to estimate the impact of volatility in the valuation of life insurance products with respect to financial markets. As future work, the reaction to the maket for an emerging market can be investigated and compared with the developed country case.

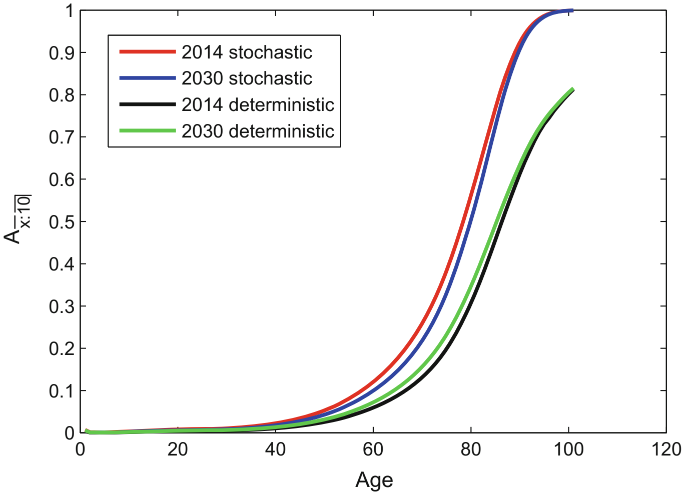


Fig. 6 Actuarial present values for a ten-year term life insurance

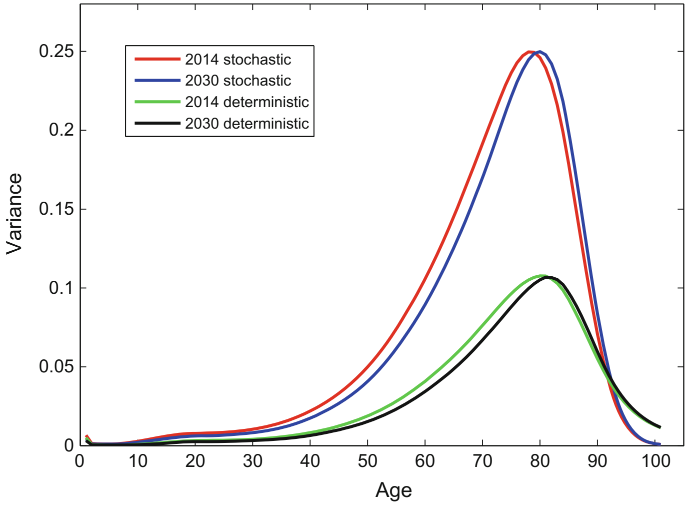


Fig. 7 Actuarial variances for a ten-year term life insurance

References

1. Boyle, P.P.: Rates of return as random variables. J. Risk Insur. **43**(4), 693–713 (1976)
2. Panjer, H.H., Bellhause, D.R.: Stochastic modeling of interest rates with applications to life contingencies. J. Risk Insur. **47**, 91–110 (1980)
3. Bellhouse, D.R., Panjer, H.: Stochastic modeling of interest rates with applications to life contingencies, Part II. J. Risk Insur. **48**, 628–637 (1981)
4. Giacotto, C.: Stochastic modeling of interest rates: actuarial versus equilibrium approach. J. Risk Insur. **53**, 435–453 (1986)
5. Dhaene, J.: Stochastic interest rates and autoregressive integrated moving average processes. ASTIN bull. **19**(2), 131–138 (1989)
6. Frees, E.W.: Net premiums in stochastic life contingencies. Trans. Soc. Actuar. **40**(1), 371–385 (1988)
7. Frees, E.W.: Stochastic life contingencies with solvency considerations. Trans. Soc. Actuar. **42**, 91–148 (1990)
8. Parker, G.: Moments of the present value of a portfolio of policies. Scand. Actuar. J. **1**, 53–67 (1994)
9. Lai, S.-W., Frees, E.W.: Examining changes in reserves using stochastic interest models. J. Risk Insur. **6**(3), 535–574 (1995)
10. Zaks, A.: Annuities under random rates of interest. Ins. Math. Econ. **28**, 1–11 (2000)
11. Debicka, J.: Moments of the cash value of future payment streams arising from life insurance contracts. Ins. Math. Econ. **33**(3), 533–550 (2003)
12. Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**, 141–183 (1973)
13. Vasicek, O.: An equilibrium characterization of the term structure. J. Financ. Econ. **5**, 177–188 (1977)
14. Cox, J.C., Ingersoll, J.E., Ross, S.A.: A theory of the term structure of interest rates. Econometrica **53**, 385–407 (1985)
15. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica **50**, 987–1007 (1982)

16. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. J. Econom. **31**, 307–327 (1986)
17. Nelson, D.B.: Conditional heteroskedasticity in asset returns: a new approach. Econometrica **59**, 347–370 (1991)
18. Ergökmen, N.G.: Stochastic modeling of random interest rates in life insurance. Unpublished M.Sc, Thesis, Middle East Technical University, Turkey (2001)
19. Market yield on U.S. Treasury securities. http://www.federalreserve.gov/releases/h15/data.htm
20. Box, G.E.P., Jenkins, G.M.: Time Series Analysis, Forecasting and Control, 2nd edn. Holden-Day, San Francisco (1976)
21. Said, D.E., Dickey, D.A.: Testing for unit roots in autoregressive moving average models of unknown order. Biometrika **71**, 599–607 (1984)
22. Lee, R.D., Carter, L.R.: Modeling and forecasting US mortality. J. Am. Stat. Assoc. **87**(419), 659–671 (1992)
23. http://www.mortality.org/cgi-bin/hmd/country.php?cntr=USA&level=1

Itô–Taylor Expansions for Systems of Stochastic Differential Equations with Applications to Stochastic Partial Differential Equations

Fikriye Yılmaz, Hacer Öz Bakan and Gerhard-Wilhelm Weber

Abstract Stochastic differential equations (SDEs) are playing a growing role in financial mathematics, actuarial sciences, physics, biology and engineering. For example, in financial mathematics, fluctuating stock prices and option prices can be modeled by SDEs. In this chapter, we focus on a numerical simulation of systems of SDEs based on the stochastic Taylor series expansions. At first, we apply the vector-valued Itô formula to the systems of SDEs, then, the stochastic Taylor formula is used to get the numerical schemes. In the case of higher dimensional stochastic processes and equations, the numerical schemes may be expensive and take more time to compute. We deal with systems with standard n-dimensional systems of SDEs having correlated Brownian motions. One the main issue is to transform the systems of SDEs with correlated Brownian motions to the ones having standard Brownian motion, and then, to apply the Itô formula to the transformed systems. As an application, we consider stochastic partial differential equations (SPDEs). We first use finite difference method to approximate the space variable. Then, by using the stochastic Taylor series expansions we obtain the discrete problem. Numerical examples are presented to show the efficiency of the approach. The chapter ends with a conclusion and an outlook to future studies.

Keywords Systems of SDEs · Itô–Taylor expansions · Correlated Brownian motions · Vector-valued Itô formula · Stochastic partial differential equations

F. Yılmaz (✉)
Department of Mathematics, Gazi University, Ankara, Turkey
e-mail: yfikriye@gmail.com

H. Öz Bakan
Department of Mathematics, Atilim University, Ankara, Turkey
e-mail: haceroz.oz@gmail.com

G.-W. Weber
Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey
e-mail: gweber@metu.edu.tr

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics II*, Springer Proceedings in Mathematics & Statistics 195,
DOI 10.1007/978-3-319-55236-1_25

1 Introduction

There has been a great interest in the simulation methods of SDEs in the fields of, e.g., finance, insurance, medicine and the modern technologies [15–17, 21, 22]. As the need to take into account of uncertainty is more and more accepted in science and the applications, SDEs are an emerging subject of interest. Stochastic Taylor expansion provides a source for the discrete-time approximation methods. One of the simplest ways to discretize the process is *Euler* method, which approximates the integrals by using the left-point rule. The *Milstein* scheme (1974), which has the order 1.0 of strong convergence, is stronger than *Euler* method. By adding further stochastic integrals, with the equations and using the stochastic Taylor expansion, more accurate schemes can be obtained.

P.E. Kloeden and E. Platen [15] have given a methodical means of deriving the Taylor series for both Stratonovich and Itô form of a SDE. The application of Itô–Taylor formula to the 1-dimensional SDEs is given explicitly in [15]. In this case, $I_{i_1,i_2,\ldots,i_k}$ represent the Itô integral, where integration is with respect to ds if $i_k = 0$, or dZ_s if $i_k = 1$ ([15], formula (2.12) on p. 169).

By recursively using the Itô formula, the obtained Taylor series can be related to a tree theory. The tree solution is given for the true solution in [3, 9]. Runge–Kutta methods having order 1.5 have been constructed in [3–5, 8, 15, 23]. In this work, in order to get numerical solutions, we shall consider Taylor schemes that converge strongly. More details about stochastic Taylor schemes can be found in [19, 24].

Vector-valued Itô calculus may involve more than one Brownian motion. Although it seems that the extension of one-variable SDEs to the multi-valued SDEs is easy, the computations of iterated Itô integrals are very expensive in the numerical approximations. But, the systems of SDEs model the important cases having several sources of randomness and correlation; these are main reasons of financial risk and instability.

Recently, there has been a great interest in survey of stochastic partial differential equations [1, 6, 7, 14, 20]. Generally, we can not find the analytical solutions of SPDEs. So that numerical methods have been getting an important issue. In this work, we first start with a simple SPDE as an application part. We show that after obtaining the discrete problem, we confront by a system of SDEs.

There are not many packages with regards to the simulations of SDEs [11]. The Maple package *stochastic* provides a symbolic manipulation for SDEs. The package may be downloaded from www.math.uni-frankfurt.de/~numerik/kloeden. Many numerical schemes may be generated by using this package. However, the computations of the iterated multi-dimensional Itô integrals are not supported. The software package *SDE lab* generated by H. Gilsing and T. Shardlow [10] is a good source, especially, for 1-dimensional SDEs.

In this work, we provide a routine, supported by MATLAB, to compute the iterated multi-dimensional Itô integrals with presence of the formulations given in [15]. We use the *Polar Marsaglia method* to generate random variables. This method, that is attributed to G. Marsaglia, is a variation of the Box-Muller method. It is based on choosing random points (x, y) in the square given by $-1 < x < 1, -1 < y < 1$. For $s = x^2 + y^2 < 1$, the following pair of normal random variables is obtained:

$$x\sqrt{\frac{-2\ln(s)}{s}}, \quad y\sqrt{\frac{-2\ln(s)}{s}}.$$

We shall call the process $\mathbb{W}_t := (W_t^1, W_t^2, \dots, W_t^n)^T$ as a correlated Brownian motion if

$$dW_t^i dW_t^j = \rho_{ij}dt \;\; (i, j = 1, 2, \dots, n),$$

for a positive semi-definite matrix $\rho = (\rho_{ij})_{1\leq i,j\leq n}$ satisfying the following conditions:

$$\rho_{ii} = 1, \text{ and } \rho_{ij} = \rho_{ji} \in [-1, 1] \;\; (i \neq j).$$

The paper is organized as follows. Multivariable Itô calculus is reviewed in Sect. 2. In Sect. 3, the systems of SDEs with standard Brownian motion and their Taylor expansions are analyzed. Then, the correlated ones are covered in Sect. 4. Section 5 includes the discretization schemes. After applying Itô–Taylor expansions to the SPDEs in Sect. 6, our numerical results are stated in Sect. 7. We provide a conclusion and give an outlook to the future studies in Sect. 8. We give some implementation details in the appendix section.

2 Multi-dimensional Itô Calculus

We consider the process $\mathbb{X}_t$ in $\mathbb{R}^d$. Let $\mathbb{Z}_t$ be a multi-dimensional Brownian motion, defined as $\mathbb{Z}_t = (Z_t^1, Z_t^2, \dots, Z_t^n)^T$. Then, the k^{th} component of the vector-valued SDE is given by

$$dX_t^k = a_k dt + \sum_{j=1}^{n} h_{kj} dZ_t^j \quad (k = 1, 2, \dots, d),$$

where $a_k(t, \mathbb{X}_t)$ and $h_{kj}(t, \mathbb{X}_t)$ are the drift and the diffusion coefficients, respectively.

We shortly put $\mathbb{A} := \mathbb{A}(t, \mathbb{X}_t)=(a_1(t, \mathbb{X}_t), \dots, a_d(t, \mathbb{X}_t))^T$, $\mathbb{X}_t = (X_t^1, \dots, X_t^d)^T$ and

$$\mathbb{H} := \mathbb{H}(t, \mathbb{X}_t) = \begin{pmatrix} h_{11}(t, \mathbb{X}_t) & \dots & h_{1n}(t, \mathbb{X}_t) \\ \vdots & \ddots & \vdots \\ h_{d1}(t, \mathbb{X}_t) & \dots & h_{dn}(t, \mathbb{X}_t) \end{pmatrix}$$

to get the following matrix formulation:

$$d\mathbb{X}_t = \mathbb{A}dt + \mathbb{H}d\mathbb{Z}_t. \tag{1}$$

3 Itô–Taylor Approximation for Standard Brownian Motions

In this section, we will assume that the Brownian motions are independent.

Equation (1) can be written in integral form as

$$\mathbb{X}_t = \mathbb{X}_{t_0} + \int_{t_0}^{t} \mathbb{A}(s, \mathbb{X}_s)ds + \int_{t_0}^{t} \mathbb{H}(s, \mathbb{X}_s)d\mathbb{Z}_s. \tag{2}$$

The second integral is called *Itô stochastic integral*, which is defined by K. Itô in 1940 [13]. This integral can be approximated by stochastic Taylor method. Before going into the numerical schemes, we recall the multi-dimensional Itô Lemma.

Lemma 1 (The multi-dimensional Itô Lemma, [18]) *Let $\mathbb{X}_t = (X_t^1, \ldots, X_t^d)^T$ be a vector-valued Itô process satisfying system of SDEs (1). Let $g : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^p$ be a given bounded function in $C^2([0, \infty) \times \mathbb{R}^d)$. Then,*

$$dg(t, \mathbb{X}_t) = \frac{\partial g}{\partial t}dt + \sum_{i=1}^{d} \frac{\partial g}{\partial x^i}(t, \mathbb{X}_t)dX_t^i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 g}{\partial x^i x^j}(t, \mathbb{X}_t)dX_t^i dX_t^j, \tag{3}$$

where $dZ_t^i dZ_t^j = \delta_{ij}dt$ and $dZ_t^i dt = dtdZ_t^i = 0$.

We consider now the k^{th} component of the system of SDEs (1):

$$dX_t^k = a_k(t, \mathbb{X}_t)dt + \sum_{j=1}^{n} h_{kj}(t, \mathbb{X}_t)dZ_t^j \quad (k = 1, 2, \ldots, d), \tag{4}$$

where $a_k(t, \mathbb{X}_t) = a_k(t, X_t^1, \ldots, X_t^d)$ and $h_{kj}(t, \mathbb{X}_t) = h_{kj}(t, X_t^1, \ldots, X_t^d)$ with the integral form of Eq. (4):

$$X_t^k = X_{t_0}^k + \int_{t_0}^{t} a_k(s, \mathbb{X}_s^k)ds + \int_{t_0}^{t} h_{kj}(s, \mathbb{X}_s^k)dZ_s^k. \tag{5}$$

We let $g(t, \mathbb{X}_t) = (g_1(t, \mathbb{X}_t)), \ldots, g_p(t, \mathbb{X}_t))^T$ and $Y_t = g(t, \mathbb{X}_t)$ to apply the Itô Lemma of Eq. (3).

Then, the component Y_t^ℓ is given by

$$dY_t^\ell = \frac{\partial g^\ell}{\partial t}dt + \sum_{i=1}^{d} \frac{\partial g^\ell}{\partial x^i}dX_t^i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 g^\ell}{\partial x^i x^j}dX_t^i dX_t^j.$$

Equivalently,

$$dY_t^\ell = \left(\frac{\partial g^\ell}{\partial t} + \sum_{i=1}^d a_i \frac{\partial g^\ell}{\partial x^i} + \frac{1}{2}\sum_{i,j=1}^d \sum_{p=1}^n h_{jp}h_{ip}\frac{\partial^2 g^\ell}{\partial x^i x^j}\right)dt + \sum_{j=1}^d \sum_{p=1}^n h_{jp}\frac{\partial g^\ell}{\partial x^j} dZ_t^p,$$

where all derivatives of g^ℓ are to be evaluated in $(t, \mathbb{X}_t)$ and Brownian motions are uncorrelated and, in here, we note that ℓ and k represent the ℓ^{th} component of the function $g(t, \mathbb{X}_t)$ and the k^{th} component of the system of SDEs of Eq. (4), respectively.

Before obtaining the Taylor series expansions, we define the following operators:

$$\mathscr{L}^0 := \frac{\partial}{\partial t} + \sum_{i=1}^d a_i \frac{\partial}{\partial x^i} + \frac{1}{2}\sum_{i,j=1}^d \sum_{p=1}^n h_{jp}h_{ip}\frac{\partial^2}{\partial x^i x^j},$$

and

$$\mathscr{L}^j := \sum_{p=1}^d h_{pj}\frac{\partial}{\partial x^p} \quad (j = 1, 2, \dots, n).$$

This allows us to express the multi-dimensional version of the Itô Lemma in a compact way:

$$Y_t^\ell = Y_{t_0}^\ell + \int_{t_0}^t \mathscr{L}^0 g^\ell ds + \sum_{j=1}^n \int_{t_0}^t \mathscr{L}^j g^\ell dZ_s^j. \tag{6}$$

In Eq. (5), we first choose $g^\ell := a_\ell(t, \mathbb{X}_t)$ $(k = \ell)$ to apply the Itô formula, and then, we let $g^\ell := h_{\ell j}(t, \mathbb{X}_t)$ to get

$$\begin{aligned} X_t^k = X_{t_0}^k &+ \int_{t_0}^t \Bigg[a_k(t_0, \mathbb{X}_{t_0}) + \int_{t_0}^s \mathscr{L}^0 a_k(\tau, \mathbb{X}_\tau)d\tau \\ &+ \sum_{l,j=1}^n \int_{t_0}^s \mathscr{L}^j a_k(\tau, \mathbb{X}_\tau)dZ_\tau^j\Bigg]ds + \sum_{j=1}^n \int_{t_0}^t \Bigg[h_{kj}(t_0, \mathbb{X}_{t_0}) \\ &+ \int_{t_0}^s \mathscr{L}^0 h_{kj}(\tau, \mathbb{X}_\tau)d\tau + \sum_{l=1}^n \int_{t_0}^s \mathscr{L}^l h_{kj}(\tau, \mathbb{X}_\tau)dZ_\tau^l\Bigg]dZ_s^j. \end{aligned} \tag{7}$$

We can continue with an application of Itô formula from Eq. (6) to the functions $\mathscr{L}^0 a_k$, $\sum_{j=1}^n \mathscr{L}^j a_k$, $\mathscr{L}^0 h_{kj}$ and $\sum_{l=1}^n \mathscr{L}^l h_{kj}$ in Eq. (7), and then, to the functions $\sum_{j,p=1}^n \mathscr{L}^j \mathscr{L}^p h_{kj}$, $\mathscr{L}^0 \mathscr{L}^0 a_k$, $\sum_{j=1}^n \mathscr{L}^0 \mathscr{L}^j a_k$, $\sum_{j,p=1}^n \mathscr{L}^j \mathscr{L}^p a_k$, $\mathscr{L}^0 \mathscr{L}^0 h_{kj}$, $\sum_{j,p=1}^n \mathscr{L}^j \mathscr{L}^0 h_{kj}$, $\sum_{j,p=1}^n \mathscr{L}^0 \mathscr{L}^j h_{kj}$, $\sum_{j=1}^n \mathscr{L}^j \mathscr{L}^0 a_k$ to obtain Itô–Taylor expansion:

$$
\begin{aligned}
X_t^k = X_{t_0}^k &+ a_k(t_0,\mathbb{X}_{t_0})I_0 + \sum_{j=1}^{n} h_{kj}(t_0,\mathbb{X}_{t_0})I_j \\
&+ \mathscr{L}^0 a_k(t_0,\mathbb{X}_{t_0})I_{00} + \sum_{j=1}^{n} \mathscr{L}^j a_k(t_0,\mathbb{X}_{t_0})I_{j0} \\
&+ \sum_{j=1}^{n} \mathscr{L}^0 h_{kj}(t_0,\mathbb{X}_{t_0})I_{0j} + \sum_{l,j=1}^{n} \mathscr{L}^l h_{kj}(t_0,\mathbb{X}_{t_0})I_{lj} \\
&+ \mathscr{L}^0\mathscr{L}^0 a_k(t_0,\mathbb{X}_{t_0})I_{000} + \sum_{j=1}^{n} \mathscr{L}^j\mathscr{L}^0 a_k(t_0,\mathbb{X}_{t_0})I_{j00} \\
&+ \sum_{j=1}^{n} \mathscr{L}^0\mathscr{L}^j a_k(t_0,\mathbb{X}_{t_0})I_{0j0} + \sum_{j,p=1}^{n} \mathscr{L}^j\mathscr{L}^p a_k(t_0,\mathbb{X}_{t_0})I_{jp0} \\
&+ \sum_{j=1}^{n} \mathscr{L}^0\mathscr{L}^0 h_{kj}(t_0,\mathbb{X}_{t_0})I_{00j} + \sum_{j,l=1}^{n} \mathscr{L}^l\mathscr{L}^0 h_{kj}(t_0,\mathbb{X}_{t_0})I_{l0j} \\
&+ \sum_{j,l=1}^{n} \mathscr{L}^0\mathscr{L}^l h_{kj}(t_0,\mathbb{X}_{t_0})I_{0lj} + \sum_{j,p,l=1}^{n} \mathscr{L}^l\mathscr{L}^p h_{kj}(t_0,\mathbb{X}_{t_0})I_{lpj} + R_t,
\end{aligned}
\tag{8}
$$

where R_t represent the remainder term which can be expressed as:

$$
\begin{aligned}
R_t = {} & I_{0000}[\mathscr{L}^0\mathscr{L}^0\mathscr{L}^0 a_k]_{t_0,t} + \sum_{j=1}^{n} I_{j000}[\mathscr{L}^j\mathscr{L}^0\mathscr{L}^0 a_k]_{t_0,t} \\
&+ \sum_{j=1}^{n} I_{0j00}[\mathscr{L}^0\mathscr{L}^j\mathscr{L}^0 a_k]_{t_0,t} + \sum_{j,p=1}^{n} I_{pj00}[\mathscr{L}^p\mathscr{L}^j\mathscr{L}^0 a_k]_{t_0,t} \\
&+ \sum_{j=1}^{n} I_{00j0}[\mathscr{L}^0\mathscr{L}^0\mathscr{L}^j a_k]_{t_0,t} + \sum_{j,p=1}^{n} I_{p0j0}[\mathscr{L}^p\mathscr{L}^0\mathscr{L}^j a_k]_{t_0,t} \\
&+ \sum_{j,p=1}^{n} I_{0jp0}[\mathscr{L}^0\mathscr{L}^j\mathscr{L}^p a_k]_{t_0,t} + \sum_{j,p,l=1}^{n} I_{ljp0}[\mathscr{L}^l\mathscr{L}^j\mathscr{L}^p a_k]_{t_0,t} \\
&+ \sum_{j=1}^{n} I_{000j}[\mathscr{L}^0\mathscr{L}^0\mathscr{L}^0 h_{kj}]_{t_0,t} + \sum_{j,p=1}^{n} I_{p00j}[\mathscr{L}^p\mathscr{L}^0\mathscr{L}^0 h_{kj}]_{t_0,t} \\
&+ \sum_{j,l=1}^{n} I_{0l0j}[\mathscr{L}^0\mathscr{L}^l\mathscr{L}^0 h_{kj}]_{t_0,t} + \sum_{j,p,l=1}^{n} I_{pl0j}[\mathscr{L}^p\mathscr{L}^l\mathscr{L}^0 h_{kj}]_{t_0,t} \\
&+ \sum_{j=1}^{n} I_{00lj}[\mathscr{L}^0\mathscr{L}^0\mathscr{L}^l h_{kj}]_{t_0,t} + \sum_{j,p,l=1}^{n} I_{p0lj}[\mathscr{L}^p\mathscr{L}^0\mathscr{L}^l h_{kj}]_{t_0,t} \\
&+ \sum_{j,p,l=1}^{n} I_{0lpj}[\mathscr{L}^0\mathscr{L}^l\mathscr{L}^p h_{kj}]_{t_0,t} + \sum_{j,p,l,r=1}^{n} I_{rlpj}[\mathscr{L}^r\mathscr{L}^l\mathscr{L}^p h_{kj}]_{t_0,t}.
\end{aligned}
\tag{9}
$$

Here, we note that the terms $I_{i_1 i_2 \dots i_k}$ in X_t^k represent the multiple Itô integrals with constant integrands, and the terms $I_{i_1 i_2 \dots i_{k+1}}[\mathscr{L}^{i_1}\mathscr{L}^{i_2}\dots\mathscr{L}^{i_k}a_k]_{t_0,t}$ or $I_{i_1 i_2 \dots i_{k+1}}$ $[\mathscr{L}^{i_1}\mathscr{L}^{i_2}\dots\mathscr{L}^{i_k}h_{kk}]_{t_0,t}$ in R_t denote the multiple Itô integrals with nonconstant integrands.

4 Systems with Correlated Brownian Motions

We assume that Brownian motions are correlated, so we use $(\mathbb{W}_t)_{t\geq 0}$ instead of $(\mathbb{Z}_t)_{t\geq 0}$ in order to point out the difference from the standard Brownian motions. Now, we consider the system [2]:

$$dX_t^k = a_k(t,\mathbb{X}_t) + \sum_1^n h_{kj}(t,\mathbb{X}_t)dW_t^j \quad (k = 1, 2, \dots, d = n), \tag{10}$$

where $dW_t^i dW_t^j = \rho_{ij}dt$.

Then, we write the correlation matrix ρ as:

$$\rho := \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix}, \quad \rho_{ij} = \rho_{ji} \in [-1, 1].$$

Here, ρ is positive semi-definite matrix which means that $\rho = \rho^T$ and

$$\sum_{i,j=1}^n \rho_{ij}x_i x_j \geq 0,$$

for all $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$.

By using basic standard Linear Algebra, one can find an $n \times n$ matrix $\mathbb{B} = (b_{ij})_{1\leq i,j\leq n}$ such that

$$\rho = \mathbb{B}\mathbb{B}^T.$$

Moreover, using Cholesky Decomposition, we can take $\mathbb{B}$ as an upper (or lower) triangular matrix.

Correlated Brownian motions can be interpreted as

$$\mathbb{W}_t = \mathbb{B}\mathbb{Z}_t,$$

where $\mathbb{W}_t = (W_t^1, \dots, W_t^n)^T$ is a n-dimensional correlated Brownian motion and the Brownian motions, W_t^i for $i = 1, 2, \dots, n$, are correlated.

In componentwise notation,

$$W_t^i = \sum_{j=1}^{n} b_{ij} Z_t^j \quad (i = 1, 2, \dots, n). \tag{11}$$

Now, after substituting Eq. (11) into (10), we obtain a system of SDEs as in Eq. (4) so that we can apply a similar procedure of argumentation.

Example 1 We consider 2-dimensional version of *weakly-coupled Ornstein–Uhlenbeck model*:

$$\begin{aligned} dX_t^1 &= \alpha_1(\theta_1 - X_t^1)dt + \sigma_1 dW_t^1, \\ dX_t^2 &= \alpha_2(\theta_2 - X_t^2)dt + \sigma_2 dW_t^2, \end{aligned} \tag{12}$$

where $dW_t^1 dW_t^2 = \rho dt$, all coefficients are real and positive, and $\rho \in (-1, 1)$.

Then, the correlation matrix ρ can be written as:

$$\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{pmatrix},$$

by Cholesky Decomposition [2].

Thus,

$$\mathbb{W}_t = \mathbb{B}\mathbb{Z}_t = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} Z_t^1 \\ Z_t^2 \end{pmatrix}.$$

So, Eq. (12) becomes

$$\begin{aligned} dX_t^1 &= \alpha_1(\theta_1 - X_t^1)dt + \sigma_1 dZ_t^1, \\ dX_t^2 &= \alpha_2(\theta_2 - X_t^2)dt + \sigma_2 \rho dZ_t^1 + \sigma_2 \sqrt{1-\rho^2} dZ_t^2. \end{aligned} \tag{13}$$

5 Discretization Schemes with Strong Taylor Approximations

We represent some numerical approximations of Itô integrals. Consider an equi-spaced discretization $t_0 \leq \tau_0 < \tau_1 < \dots < \tau_\nu < \dots < \tau_m = T$ of the time interval $[t_0, T]$. Let $\Delta = T/m$ denote the increments (step-size); then, for all $\nu \in \{0, 1, \dots, m-1\}$ it holds:

$$\begin{aligned}
I_j &= \int_{\tau_\nu}^{\tau_{\nu+1}} dZ_s^j = \Delta Z^j = Z_{\tau_{\nu+1}}^j - Z_{\tau_\nu}^j, \\
I_{j0} &= \int_{\tau_\nu}^{\tau_{\nu+1}} \int_{\tau_\nu}^{s} dZ_u^j ds = \Delta \tilde{W}, \\
I_{0j} &= \int_{\tau_\nu}^{\tau_{\nu+1}} \int_{\tau_n}^{s} du dZ_s^j = (\Delta Z^j)\Delta - \Delta \tilde{W}, \\
I_{jj} &= \frac{1}{2}\big((\Delta Z^j)^2 - \Delta\big), \\
I_{jj0} &= I_{0jj} = I_{j0j} = \frac{1}{6}\Delta\big((\Delta Z^j)^2 - \Delta\big), \\
I_{jjj} &= \frac{1}{6}\Delta\big((\Delta Z^j)^3 - 3\Delta(\Delta Z^j)\big), \\
I_{j00} &= I_{0j0} = I_{00j} = \frac{1}{6}\Delta^2 \Delta Z_s^j,
\end{aligned}$$

where the $\Delta \tilde{W}$ and ΔZ^j are Gaussian random variables with $\Delta Z^j \sim N(0, \Delta)$, $\Delta \tilde{W} \sim N(0, \frac{1}{3}\Delta^3)$ and $E\big(\Delta Z^j \Delta \tilde{W}\big) = \frac{1}{2}\Delta^2$.

Now, we shall use the above relations to propose some strong approximations. Firstly, let us consider the Order 1.5 Strong Taylor Scheme. The k^{th} component of *the order 1.5 strong Taylor scheme* is given by

$$\begin{aligned}
X_{\nu+1}^k = X_\nu^k &+ a_k\Delta + \frac{1}{2}\mathscr{L}^0 a_k \Delta^2 + \sum_{j=1}^{n} (h_{kj}\Delta Z^j + \mathscr{L}^0 h_{kj} I_{0j} + \mathscr{L}^j a_k I_{j0}) \\
&+ \sum_{j_1, j_2 = 1}^{n} \mathscr{L}^{j_1} h_{kj_2} I_{j_1 j_2} + \sum_{j_1, j_2, j_3 = 1}^{n} \mathscr{L}^{j_1} \mathscr{L}^{j_2} h_{kj_3} I_{j_1 j_2 j_3} \qquad (14)
\end{aligned}$$

$(\nu = 0, 1, 2, \dots, m - 1)$.

Similarly, the k^{th} component of *the order 2.0 strong Taylor scheme* takes the form

$$\begin{aligned}
X_{\nu+1}^k = X_\nu^k &+ a_k\Delta + \frac{1}{2}\mathscr{L}^0 a_k \Delta^2 + \sum_{j=1}^{n} (h_{kj}\Delta Z^j + \mathscr{L}^0 h_{kj} I_{0j} + \mathscr{L}^j a_k I_{j0}) \\
&+ \sum_{j_1, j_2 = 1}^{n} \big(\mathscr{L}^{j_1} h_{kj_2} I_{j_1 j_2} + \mathscr{L}^0 \mathscr{L}^{j_1} h_{kj_2} I_{0 j_1 j_2} + \mathscr{L}^{j_1} \mathscr{L}^0 h_{kj_2} I_{j_1 0 j_2} \\
&+ \mathscr{L}^{j_1} \mathscr{L}^{j_2} a_k I_{j_1 j_2 0}\big) + \sum_{j_1, j_2, j_3 = 1}^{n} \mathscr{L}^{j_1} \mathscr{L}^{j_2} h_{kj_3} I_{j_1 j_2 j_3} \qquad (15) \\
&+ \sum_{j_1, j_2, j_3, j_4 = 1}^{n} \mathscr{L}^{j_1} \mathscr{L}^{j_2} \mathscr{L}^{j_2} h_{kj_4} I_{j_1 j_2 j_3 j_4}
\end{aligned}$$

$(\nu = 0, 1, 2, \dots, m - 1)$.

6 Application to Stochastic Partial Differential Equations

SPDEs have also become quite popular models in financial mathematics. For example, they are used to describe the investment performance process in portfolio choice models; maximizing the expected utility of terminal wealth, so-called forward investment performance criterion [14, 25]. To analyze SPDEs help us to get a better understanding of the characteristics of the value function and optimal investment decision.

In $\mathbb{R}^d$, we consider the following SPDEs

$$d\mathbb{X}_t = (A\mathbb{X}_t + f(\mathbb{X}_t))dt + g(\mathbb{X}_t)dZ_t, \tag{16}$$

where A represents the partial differential operator and we suppose the coefficients satisfy all the criterion for existence and uniqueness of the solution.

In order to discretize the continuous SPDE given in Eq. (16), we consider the corresponding Itô–Galerkin SDE in $\mathbb{R}^N$, with some $N \in \mathbb{N}$ [14]. Then,

$$d\mathbb{X}_t = (A\mathbb{X}_t + f(\mathbb{X}_t))dt + g(\mathbb{X}_t)dZ_t$$

is an N-dimensional SDE in $\mathbb{R}^N$. We consider the one dimensional SPDE of the form

$$\begin{cases} du(t,x) = \left(\frac{\partial^2}{\partial x^2}u(t,x) - bu(t,x)\right)dt + f(t,x)dZ_t, \quad t \geq 0, \\ u(0,x) = u_0(x), \quad 0 \leq x \leq 1, \\ u(t,0) = u(t,1) = 0, \quad t \geq 0. \end{cases}$$

While we apply the Itô–Taylor method to the time variable, we use a finite difference scheme to approximate the space variable.

Example 2 We consider the 1-dimensional semilinear stochastic heat equation of the form

$$\begin{cases} du(t,x) = \left(\frac{\partial^2}{\partial x^2}u(t,x) + \frac{1}{2}u(t,x)\right)dt + u(t,x)dZ_t, \quad t \geq 0, \\ u(t,0) = u(t,1) = 0, \quad t \geq 0, \\ u_0(x) = \sin(\pi x), \quad 0 \leq x \leq 1. \end{cases}$$

For space variable, $0 \leq x_0 < x_1 < \ldots < x_j < \ldots < x_M = 1$ denotes the equispaced discretization of time space interval $[0, 1]$. Then, by applying the second-order central difference scheme for the space variable, we get

$$du_t^j = \left(\frac{u_t^{j-1} - 2u_t^j + u_t^{j+1}}{h^2} + \frac{1}{2}u_t^j\right)dt + u_t^j dZ_t,$$

where

$$h = x_{j+1} - x_j = \frac{1}{M} \quad (j = 0, 1, \dots, M-1).$$

Then, we obtain a system of equations for $j = 1, 2, \dots, M$

$$\begin{aligned}
du_t^1 &= \left(\frac{u_t^0 + u_t^2}{h^2} + \lambda u_t^1\right) dt + u_t^1 dZ_t, \\
du_t^2 &= \left(\frac{u_t^1 + u_t^3}{h^2} + \lambda u_t^2\right) dt + u_t^2 dZ_t, \\
&\vdots \\
du_t^j &= \left(\frac{u_t^{j-1} + u_t^{j+1}}{h^2} + \lambda u_t^j\right) dt + u_t^j dZ_t, \\
&\vdots \\
du_t^M &= \left(\frac{u_t^{M-1} + u_t^{M+1}}{h^2} + \lambda u_t^M\right) dt + u_t^M dZ_t,
\end{aligned}$$

where

$$\lambda = \left(\frac{1}{2} - \frac{2}{h^2}\right).$$

In matrix notation, we have

$$d\mathbb{U}_t = (\mathbb{A}_t + \mathbb{B}_t)dt + d\mathbb{Z}_t,$$

where

$$\mathbb{A}_t = \begin{pmatrix} \lambda & 1/h^2 & 0 & \cdots \\ 1/h^2 & \lambda & 1/h^2 & \cdots \\ 0 & 1/h^2 & \lambda & \cdots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

$\mathbb{U}_t = (u_t^1, u_t^2, \dots, u_t^M)^T$, $\mathbb{B}_t = (u_t^0/h^2, 0, \dots, 0)^T$ and $\mathbb{Z}_t = (dZ_t, \dots, dZ_t)^T$.

The j^{th} component of the Euler scheme can be obtained as follows:

$$u_t^j = u_0^j + \left(\frac{u_0^{j-1} + u_0^{j+1}}{h^2} + \lambda u_0^j\right) \Delta + u_0^j \Delta Z \quad (j = 1, 2, \dots, M).$$

Moreover, the j^{th} component of the Milstein scheme can be obtained as follows:

$$u_t^j = u_0^j + \left(\frac{u_0^{j-1} + u_0^{j+1}}{h^2} + \lambda u_0^j \right) \Delta + u_0^j \Delta Z + \frac{1}{2} u_0^j ((\Delta Z)^2 - \Delta)$$

$(j = 1, 2, \ldots, M)$.

7 Numerical Results

In this section, we consider numerical examples for both the systems with independent and correlated Brownian motions. In order to implement the discrete scheme, we use MATLAB. There are many documentations that describes the main features of MATLAB commands related to SDE. Some numerical interpretations can be found in the SDEs MATLAB packages [11]. However, numerical examples implemented in MATLAB are mostly in the 1-dimensional case. Some coupled SDEs are considered, but having symmetric coefficients allowing easy computations that arise from multiple Itô integrals. Our first example demonstrates the triple SDEs having independent Brownian motions.

Example Run 1. The system of SDE consisting of three equations proposed in Hofmann, Platen and Schweizer [12] is considered as:

$$\begin{cases} dX_t^1 = X_t^1 X_t^2 dZ_t^1, \\ dX_t^2 = -(X_t^2 - X_t^3)dt + 0.3 X_t^2 dZ_t^2, \\ dX_t^3 = \frac{1}{\alpha}(X_t^2 - X_t^3)dt, \end{cases}$$

where X_t^1, X_t^2 and X_t^3 represent the asset price, the instantaneous volatility, and the averaged volatility, respectively, and, Z_t^1 and Z_t^2 are uncorrelated Brownian motions. As in [11], Milstein scheme is obtained as:

$$\begin{aligned} X_{\nu+1}^1 &= X_\nu^1 + X_\nu^1 X_\nu^2 \Delta Z^1 + \frac{1}{2} X_\nu^1 (X_\nu^2)^2 \{(\Delta Z^1)^2 - \Delta\} \\ &\qquad + 0.3 X_\nu^1 X_\nu^2 \int_{t_\nu}^{t_{\nu+1}} \int_{t_\nu}^{t} dZ_s^2 dZ_s^1, \\ X_{\nu+1}^2 &= X_\nu^2 - (X_\nu^2 - X_\nu^3)\Delta + 0.3 X_\nu^2 \Delta Z^2 + 0.045 X_\nu^2 \{(\Delta Z^1)^2 - \Delta\}, \\ X_{\nu+1}^3 &= X_\nu^3 + \frac{1}{\alpha}(X_\nu^2 - X_\nu^3)\Delta \quad (\nu \in \mathbb{N}). \end{aligned}$$

We take $\alpha = 1$, $X_0^1 = 1$, $X_0^2 = 0.1$, $X_0^2 = 0.1$ and $T = 1$; Δ is considered as 2^{-9}.

The scheme has the double integral $\int_{t_\nu}^{t_{\nu+1}} \int_{t_\nu}^{t} dZ_s^2 dZ_s^1$. In [11], this integral is approximated by Euler method. Although it is a bit challenging, we compute such integrals by using the following formulations from [15] as follows:

$$
\begin{aligned}
&I_0^p = \Delta, \quad I_j^p = \sqrt{\Delta}\xi_j, \quad I_{00}^p = \frac{1}{2}\Delta^2, \\
&I_{j0}^p = \frac{1}{2}\Delta\big(\sqrt{\Delta}\xi_j + a_{j0}\big), \quad I_{0j}^p = \frac{1}{2}\Delta\big(\sqrt{\Delta}\xi_j - a_{j0}\big).
\end{aligned}
$$

Here,

$$
\begin{aligned}
a_{j0} &= -\frac{1}{\pi}\sqrt{2\Delta}\sum_{r=1}^{p}\frac{1}{r}\zeta_{jr} - 2\sqrt{\Delta\rho_p}\mu_{jp}, \\
I_{j_1 j_2}^p &= \frac{1}{2}\Delta\xi_{j_1}\xi_{j_2} - \frac{1}{2}\sqrt{\Delta}(a_{j_2 0}\xi_{j_1} - a_{j_1 0}\xi_{j_2}) + \Delta A_{j_1 j_2}^p, \\
A_{j_1 j_2}^p &= \frac{1}{2\pi}\sum_{r=1}^{p}\frac{1}{r}(\zeta_{j_1 r}\eta_{j_2 r} - \zeta_{j_2 r}\eta_{j_1 r}),
\end{aligned}
$$

with

$$
\begin{aligned}
&\xi_j = \frac{1}{\sqrt{\Delta}}W^j, \quad \zeta_{jr} = \sqrt{\frac{2}{\Delta}}\pi r a_{jr}, \quad \eta_{jr} = \sqrt{\frac{2}{\Delta}}\pi r b_{jr}, \\
&\mu_{jp} = \frac{1}{\sqrt{\Delta\rho_p}}\sum_{r=p+1}^{\infty} a_{jr}, \quad \rho_p = \frac{1}{12} - \frac{1}{2\pi^2}\sum_{r=1}^{p}\frac{1}{r^2},
\end{aligned}
$$

where $j = 1, 2, \dots, m$ and $r = 1, 2, \dots, p$, for a positive number p chosen as follows:

$$
p = p(\Delta) \geq \frac{K}{\Delta^2},
$$

for an appropriate positive constant K to ensure the convergence order of the numerical scheme.

Here, we note that ζ_{jr}, η_{jr} and μ_{jp} are uncorrelated Gaussian random variables. We use the *Polar Marsaglia Method* to generate the pairs of random variables. Implementation of this method in MATLAB can be found in the Appendix. Moreover, the codes of the computation of iterated integral $\int_{\nu_r}^{t_{\nu+1}} \int_{\nu_r}^{t} dZ_s^2 dZ_s^1$ can also be found in the Appendix.

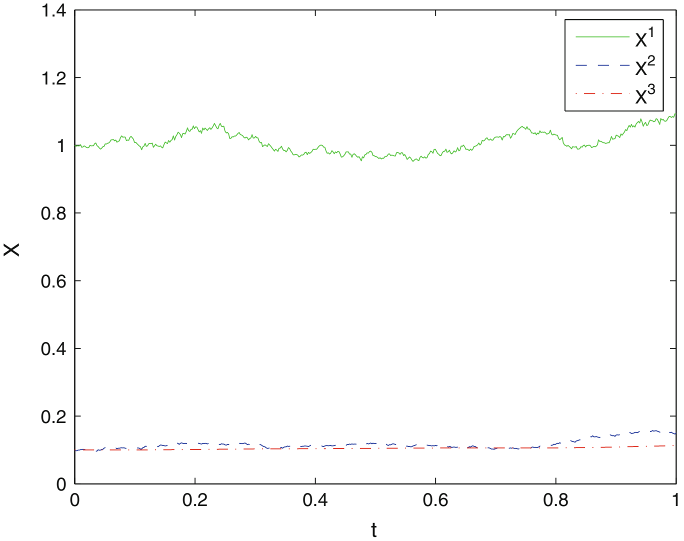


Fig. 1 Numerical result of Example Run1 with Milstein approximation

In Fig. 1, we give the numerical result of Example Run1. We have the same results as in [11].

Example Run 2. (*Correlated Brownian motions*) We consider the strongly-coupled Ornstein–Uhlenbeck process:

$$\begin{aligned} dX_t^1 &= (-\alpha_{11}X_t^1 - \alpha_{12}X_t^2)dt + \sigma_1 dW_t^1, \\ dX_t^2 &= (-\alpha_{21}X_t^1 - \alpha_{22}X_t^2)dt + \sigma_2 dW_t^2, \end{aligned} \tag{17}$$

where $dW_t^1 dW_t^2 = \rho dt$ and the transformed form of Eq. (17) is

$$\begin{aligned} dX_t^1 &= (-X_t^1 - 2X_t^2)dt + dZ_t^1, \\ dX_t^2 &= (-X_t^1 - X_t^2)dt + 0.6dZ_t^1 + 0.8dZ_t^2. \end{aligned}$$

Our Taylor scheme with order 1.5 gives

$$\begin{aligned} X_{\nu+1}^1 &= X_\nu^1 + (-X_\nu^2 - 2X_\nu^1)\Delta + \frac{1}{2}(3X_\nu^2 + 5X_\nu^1)\Delta^2 + \Delta Z^1 - 2.6I_{10} - 0.8I_{20}, \\ X_{\nu+1}^2 &= X_\nu^2 + (-X_\nu^2 - X_\nu^1)\Delta + \frac{1}{2}(2X_\nu^2 + 3X_\nu^1)\Delta^2 + 0.6\Delta Z^1 \\ &\qquad + 0.8\Delta Z^2 - 1.6I_{10} - 0.8I_{20} \quad (\nu \in \mathbb{N}). \end{aligned}$$

For the computation of integrals I_{20} and I_{10} and the numerical simulation of the example, see the Appendix.

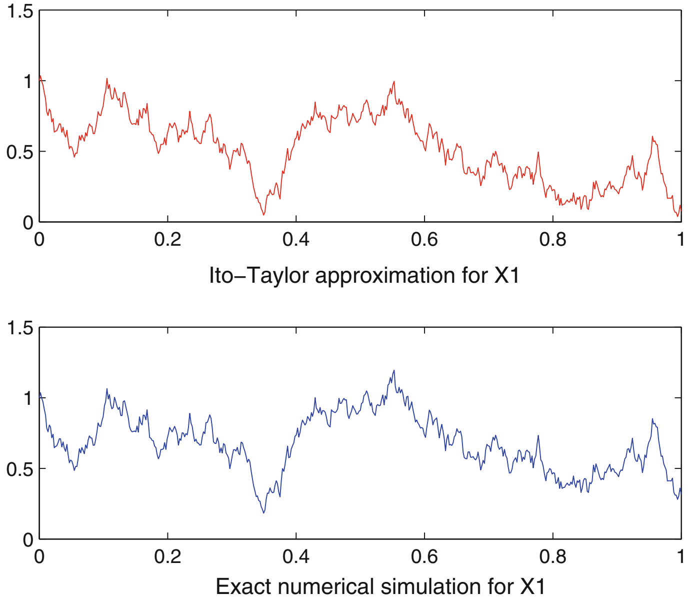


Fig. 2 Comparison of numerical simulation for the analytical solution of Run2 with Taylor Scheme of order 1/2 for X_1

The exact solution of the system from Eq. (17) is obtained in matrix formulation as:

$$\mathbb{X}_t = \mathbb{X}_0 \exp(-t\mathbb{A}) + \int_0^t \exp((s-t)\mathbb{A})\mathbb{B}d\mathbb{Z}_t, \tag{18}$$

where $\mathbb{A} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$ and $\mathbb{B} = \begin{pmatrix} 1 & 0 \\ 0.6 & 0.8 \end{pmatrix}$.

We perform the numerical simulation for the analytical solution of Eq. (18). We first compute the matrix multiplications in Eq. (18), and then approximate componentwise. In Fig. 2, we compare the obtained results for X1. It can be seen easily that they are almost the same.

Example Run 3. As we discussed before, we consider the 1-dimensional semilinear stochastic heat equation of the form

$$\begin{cases} du(t,x) = \left(\frac{\partial^2}{\partial x^2}u(t,x) + \frac{1}{2}u(t,x)\right)dt + u(t,x)dZ_t, \quad t \geq 0, \\ u(t,0) = u(t,1) = 0, \quad t \geq 0, \\ u_0(x) = \sin(\pi x), \quad 0 \leq x \leq 1. \end{cases}$$

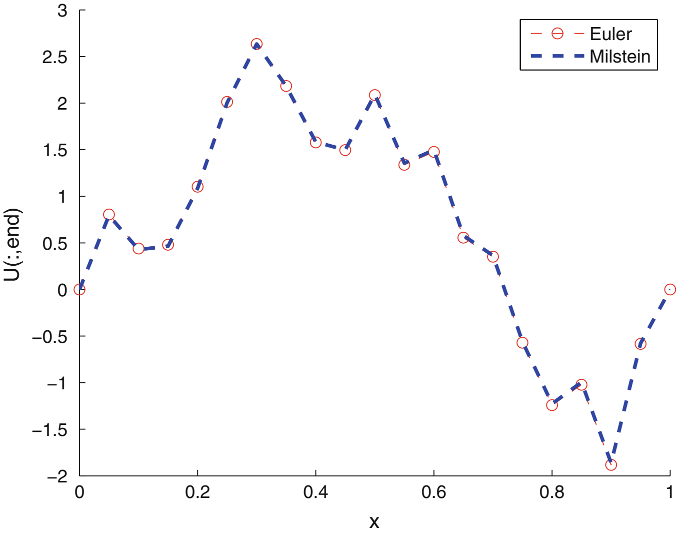


Fig. 3 Comparison of Backward Euler and Milstein Schemes for Run 3

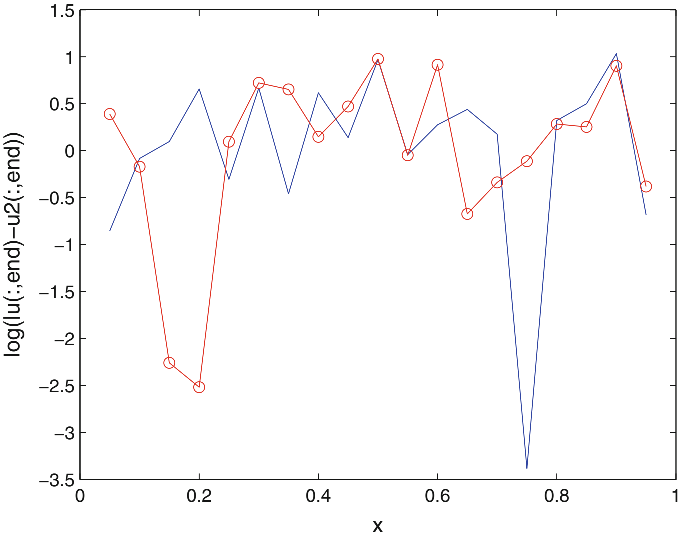


Fig. 4 Log difference numerical approximation to solutions with $\Delta = 1/1000$ (*blue*) and $\Delta = 1/3000$ (*red*) compared to the fixed solution

We recall that the j^{th} component of the Milstein scheme is:

$$u_{\nu+1}^{j} = u_{\nu}^{j} + \left(\frac{u_{\nu}^{j-1} + u_{\nu}^{j+1}}{h^2} + \lambda u_{\nu}^{j} \right) \Delta + u_{\nu}^{j} \Delta Z + \frac{1}{2} u_{\nu}^{j} ((\Delta Z)^2 - \Delta) \quad (\nu \in \mathbb{N}).$$

As seen in Fig. 3, our method with Milstein scheme and the Backward Euler method gives almost the same result.

Since we do not have the exact solution for this example, we fix $h = 1/20$ and $\Delta = 1/2000$ by preserving the numerical stability of the stochastic heat equation. Then, we compare the logarithmic difference between the fixed solution and solutions corresponding to $\Delta = 1/1000$ and $\Delta = 1/3000$ in Fig. 4.

8 Conclusion

In this chapter, we tried to display a part of the inner beauty of stochastic dynamics and, herewith, of various kinds of optimization and optimal control problems subject to those dynamical constraints. Here, this beauty expresses itself in terms of *digitalization*, *algebraization* and *automization* which are not only very aesthetic indeed, but also very practical.

Future research in this field can fruitfully address various extensions and many real-world applications in the areas of finance, but also actuarial sciences, microbiology, communication and social sciences, and the wide field of modern economics.

Moreover, *digitalization* of expectations can be considered for the order issues of higher-order weak schemes. As a mathematical application, control problems of SDEs may be worked out, further along the lines of this chapter.

Appendix

Polar Marsaglia Method:

```
%Polar Marsaglia Method
function [z1,z2]= Polar
l=0.5;
while l>0
u1 = rand;u2 = rand;
v1 = 2*u1 - 1;v2 = 2*u2 - 1;
V = (v1.*v1)+(v2.*v2);
if (V<=1)&&(V>0)
    break;
end
```

```
end
z1 = v1.*sqrt(-2*log(V)./V);
z2 = v2.*sqrt(-2*log(V)./V).
```

Approximation of I_{ij}:

```
%Approximation of I_ij
function I_ij=ito_ij(p,Delta,G1,G2,mu1_j,mu2_j,ro)
a_ij=0;
for i=1:p
    [zeta1, zeta2]= Polar;[eta1 ,eta2 ]=Polar;
    a_ij=a_ij+(1/i)*(zeta1*(sqrt(2)*G2+eta2)...
    -zeta2*(sqrt(2)*G1+eta1));
end
I_ij=a_ij*Delta/(pi);
I_ij=I_ij+Delta*(G1*G2/2+sqrt(ro)*(mu1_j*G2-mu2_j*G1)).
```

Approximation of I_{j0} and I_{0j}:

```
%Approximation of I_j0 and I_0j
function [I_10,I_20]=ito_j0(p,Delta,G1,G2,mu1_j,mu2_j,ro)
a_10=0;a_20=0;
for i=1:p
   [eta1,eta2]=Polar;
    a_10=a_10+(1/i)*eta1;
    a_20=a_20+(1/i)*eta2;
end
I_10=a_10*(1/pi)*sqrt(Delta*2)+2*sqrt(Delta*ro)*mu1_j;
I_10=(1/2)*Delta*I_10+(1/2)*Delta*sqrt(Delta)*G1;
I_20=a_20*(1/pi)*sqrt(Delta*2)+2*sqrt(Delta*ro)*mu2_j;
I_20=(1/2)*Delta*I_20+(1/2)*Delta*sqrt(Delta)*G2;
```

The main file can be run as:

```
clf
randn(’state’,1)
T = 1; Delta = 2^(-9); delta = Delta^2;
L = T/Delta; K = Delta/delta;
X1 = zeros(1,L+1); X2 = zeros(1,L+1);
X1(1) = 1;X2(1) = 0.1;
p=2;ro=0;
for i=1:p
   ro=ro+1/(i*i);ro=(pi*pi)/6-ro;ro=ro/(2*pi*pi);
end
for j = 1:L
```

```
    G1 = randn;   G2 = randn;
Winc2 = sqrt(Delta)*G2;Winc1 = sqrt(Delta)*G1;
 [mu1, mu2 ]=Polar;

[I10,I20]=ito_j0(p,Delta,G1,G2,mu1,mu2,ro);
X1(j+1) = X1(j) +(-2*X1(j)-X2(j))*Delta +...
    (0.5)*(3*X2(j)+5*X1(j))*Delta^2+Winc1-...
    (2.6)*I10-(0.8)*I20;
X2(j+1) = X2(j) + (-X2(j)-X1(j))*Delta+(0.5)*(2*X2(j)...
    +3*X1(j))*Delta^2+ (0.6)*Winc1+...
    (0.8)*Winc2-(1.6)*I10-(0.8)*I20;
end
plot([0:Delta:T],X1,'r-'), hold on
plot([0:Delta:T],X2,'bl--')
xlabel('t','FontSize',16), ylabel('X','FontSize',16)
legend('X^1','X^2')
```

References

1. Barth, A., Lang, A.: Simulation of stochastic partial differential equations using finite element methods. Stoch. Int. J. Probab. Stoch. Process. **84**(2–3), 217–231 (2012)
2. Brummelhuis, R.: Mathematical Methods Lecture Notes 6, Department of Economics, Mathematics and Statistics, Birkbeck, University of London (2009)
3. Burrage, K., Burrage, P.M.: High strong order methods for non-commutative stochastic ordinary differential equation system and the Magnus formula. Physica D **133**, 34–48 (1999)
4. Burrage, K., Burrage, P.M.: Order conditions of stochastic Runge-Kutta methods by B-series. SIAM J. Numer. Anal. **38**, 1626–1646 (1999)
5. Burrage, P.M.: Runge-Kutta Methods for Stochastic Differential Equations, Ph.D. Thesis, Department of Mathematics, University of Queensland, Australia (1999)
6. Davie, A.M., Gaines, J.G.: Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations. Math. Comp. **70**, 121–134 (2001)
7. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions, Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge (1992)
8. Debrabant, K., Kværnø, A.: B-series analysis of stochastic Runge-Kutta methods that use an iterative scheme to compute their internal stage values. SIAM J. Numer. Anal. **47**(1), 181–203 (2008/09)
9. Debrabant, K., Kværnø, A.: Composition of stochastic B-series with applications to implicit Taylor methods. Appl. Numer. Math. **61**(4), 501–511 (2011)
10. Gilsing, H., Shardlow, T.: SDELab: a package for solving stochastic differential equations in MATLAB. J. Comput. Appl. Math. **205**(2), 1002–1018 (2007)
11. Higham, D.J., Kloeden, P.E.: MAPLE and MATLAB for stochastic differential equations in finance. In: Programming Languages and Systems in Computational Economics and Finance, pp. 233–270 (2002)
12. Hofmann, N., Platen, E., Schweizer, M.: Option pricing under incompleteness and stochastic volatility. J. Math. Financ. **2**, 153–187 (1992)
13. Ito, K.: Stochastic integral. Proc. Imperial Acad. Tokyo **20**, 519–524 (1944)
14. Jentzen, A., Kloeden, P.E.: The numerical approximation of stochastic partial differential equations. Milan J. Math. **77**, 205–244 (2009)

15. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations, vol. 21, 2nd edn. Springer, Berlin (1999)
16. Kunita, H.: Stochastic Flows and Stochastic Differential Equations. Cambridge Studies in Advanced Mathematics, vol. 24. Cambridge University Press, Cambridge (1990)
17. Milstein, G.N.: Numerical Integration of Stochastic Differential Equations. Kluwer, Dordrecht (1995)
18. Øksendal, B.: Stochastic Differential Equations, An Introduction with Applications, 5th edn. Springer, Berlin (2000)
19. Öz, H.: Advances and Applications of Stochastic Itô-Taylor Approximation and Change of Time Method: in the Financial Sector, MSc. Thesis, METU (2013)
20. Pardoux, É.: Two-sided stochastic calculus for SPDEs. Stochastic partial differential equations and applications (Trento, 1985). Lecture Notes in Math, vol. 1236, p. 200207. Springer, Berlin (1987)
21. Platen, E.: An introduction to numerical methods for stochastic differential equations. Acta Numerica **8**, 197–246 (1999)
22. Rümelin, W.: Numerical treatment of stochastic differential equations. SIAM J. Numer. Anal. **19**, 604–613 (1982)
23. Rößler, A.: Runge-Kutta methods for the strong approximation of solutions of stochastic differential equations. SIAM J. Numer. Anal. **48**(3), 922–952 (2010)
24. Yılmaz, F., Öz, H., Weber, G.W.: Calculus and “Digitalization" in Finance: Change of time method and stochastic Taylor expansion with computation of expectation. In: Pinto, A.A., Zilberman, D. (eds.) Modeling, Dynamics, Optimization and Bioeconomics I. Springer Proceedings in Mathematics Statistics, vol. 73, pp. 739–753. Springer International Publishing, Berlin (2014). ISBN 978-3-319-04848-2
25. Zaripopoulou, T., Musiela, M.: Stochastic partial differential equations and portfolio choice. In: Chiarella, C., Novikov, A. (eds.) Contemporary Quantitative Finance, pp. 195–215. Springer, Berlin (2010)