Springer Proceedings in Mathematics & Statistics

Alberto A. Pinto
David Zilberman *Editors*

Modeling, Dynamics, Optimization and Bioeconomics III

DGS IV, Madrid, Spain, June 2016, and Bioeconomy VIII, Berkeley, USA, April 2015—Selected Contributions

Springer

Springer Proceedings in Mathematics & Statistics

Volume 224

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 III

DGS IV, Madrid, Spain, June 2016, and Bioeconomy VIII, Berkeley, USA, April 2015—Selected Contributions

Springer

Editors
Alberto A. Pinto
LIAAD—INESC TEC, Department of Mathematics, Faculty of Science
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 978-3-319-74085-0 ISBN 978-3-319-74086-7 (eBook)
https://doi.org/10.1007/978-3-319-74086-7

Library of Congress Control Number: 2018930243

Mathematics Subject Classification (2010): 37-XX, 49-XX, 91-XX, 58-XX, 60-XX, 62-XX, 97M10, 97M40

Printed on acid-free paper

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

Foreword

L'esprit n'use de sa faculté créatrice que quand l'expérience lui en impose la nécessité.

Henri Poincaré

The tremendous challenges that are faced by humanity as the 21st century unfolds itself require a multidisciplinary approach. Now, more than ever before, the need for exploring our creative capacities to solve relevant problems for society is crucial for our very survival. However, this can only be done by a cross-cultural and multidisciplinary networking effort.

Having a group of scientists from different areas networking and exchanging experiences in a vibrant environment requires the correct environment. One can safely say that such environment came about in the two opportunities connected to the present volume of papers, namely in Berkeley, at the University of California, during the month of March 2014, and in Madrid, at the Universidad Nacional de Educatión a Distancia (UNED), during the month of June 2016.

In the first occasion, during the Seventh Berkeley Bioeconomy Conference, the ideal environment of the San Francisco Bay Area, with its sunny days and foggy afternoons, conjoined with the highly inquisitive and revolutionary tradition of the Cal Berkeley Campus had as its central theme "Biofuels as part of a sustainable strategy". In this occasion, an array of leading experts under the coordination of David Zilberman tackled topics ranging from global biofuel investments to the future of Brazilian biofuels, passing through extreme weather, biotechnology and agricultural productivity. In the present volume, the paper "Simulation and Advanced Control of the Continuous Biodiesel Production Process" by Brásio et al. and "Myopia of Governments and Optimality of Irreversible Pollution Accumulation" by Policardo are good examples of a quantitative follow-up of the conference themes. The paper by Mendes et al. deals with modelling by differential equations the kinetic separation of hexane isomers when they flow through a packed bed containing the micro-porous Metal-Organic Framework (MOF) ZIF-8 adsorbent. It is shown that a proper combination of two characteristic times can lead

to very different dynamics of fixed bed adsorbers wherein a limiting case can give rise to a spontaneous breakthrough curve of solutes.

In the second occasion, in the quiet Madrileño Summer and under the auspices of the UNED, we had the "4th International Conference on Dynamics, Games and Science" with the key topic being decision models in a complex economy. Here, under the warm hospitality of our Spanish colleagues, I was pleased to witness a broad plethora of distinguished speakers discussing topics ranging from human decisions, from a game theoretical viewpoint, to the simulation of energy demand and efficiency and passing through swarms of interacting agents in random environments. The remaining papers that can be found in the present volume are in a certain sense a written testimony of such diversity and effusiveness of interactions. For instance, the article by dos Santos et al. studies the influence of human mobility of dengue's transmission in the state of Rio de Janeiro from a statistical viewpoint. Still within the area of statistics, but from a broad theoretical perspective, the chapter by Casaca discusses prior information in Bayesian linear multivariate regression. The work of Nassif et al. presents a mathematical model for the tick life cycle based on the McKendrick partial differential equation. The article of Balsa et al. proposes a two-phase acceleration technique for the solution of symmetric and positive-definite linear systems with multiple right-hand sides. The paper by Rüppel et al. presents a constructive proof of the complete nonholonomy of the rolling ellipsoid. The two articles by Lopes and collaborators deal with important theoretical aspects of dynamical systems (such as the fat attractor) and quantum mechanics.

Summing up, the present volume displays a variety of works by leading researchers in a broad range of subjects where mathematical models have a substantial role and impact in society.

Rio de Janeiro, Brasil Jorge P. Zubelli
March 2017

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 Jorge P. Zubelli 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 European Regional Development Fund (ERDF) through the COMPETE Program (operational programme 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/N0RTE-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).

Contents

Optimal Regional Regulation of Animal Waste

Antti Iho, Doug Parker and David Zilberman

Abstract Large animal facilities generate manure in excess of their production needs leading to excessive nutrient loading. Differences in manure contents of phosphorus and nitrogen relative to crop requirements exacerbate loading of the more abundant nutrient, frequently phosphorus. Current regulations that restrict manure utilization and animal production, but not at crop lands leads to suboptimal resource allocation and under utilization of manure in crop production. The transboundary character of nutrient loading further complicates the management of manure phosphorus and nitrogen. Due to differences in environmental characteristics, upstream and downstream regions may have differing objectives towards controlling nitrogen and phosphorus surpluses. We consider optimal management of manure in a stylized two-agent, two-nutrient and two-region model. We show that trade-offs in managing manure phosphorus and nitrogen, inability to regulate manure applications outside animal farms' field areas and regional differences in environmental targets can severely impede the effectiveness of regulation. Depending on the environmental and economic characteristics, tightening upstream regulation with respect to the loading of one nutrient might increase the downstream loading of the other and might even decrease the total welfare.

Keywords Manure · Phosphorus · Nitrogen · Regulation · Externality
Transboundary pollution

A. Iho (✉)
Natural Resources Institute Finland (Luke), Latokartanonkaari 9,
00790 Helsinki, Finland
e-mail: antti.iho@luke.fi

D. Parker
California Institute for Water Resources, UC Agriculture and Natural Resources, 1111 Franklin Street, Oakland, CA 94670, USA
e-mail: Doug.Parker@ucop.edu

D. Zilberman
Department of Agricultural and Resource Economics, University of California Berkeley, 207 Giannini Hall 3310, Berkeley, CA 94720-3310, USA
e-mail: zilber11@berkeley.edu

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_1

1 Introduction

Animal waste is a poster child of the complexity of environmental regulation. Manure is generated in animal farms as a by-product. Its storage and application to fields as a fertilizer are linked to variety of air and water quality problems: odor, emissions of fine particulates and greenhouse gases, elevated nitrate levels in groundwater aquifers and eutrophication of surface waters [1].

Because manure influences a range of environmental attributes via different media, there are potential problems in focusing on a single problem at the time. Aillery et al. [1], for instance, analyzes tradeoffs in manure nitrogen management in water and air protection. For mitigating eutrophication of surface waters, the trade-off between nitrogen and phosphorus surpluses becomes important. For a given animal and diet, nitrogen and phosphorus in manure come in fixed proportions. The proportions of agronomic requirements of crops for nitrogen and phosphorus are also approximately fixed. Typically, manure applications are balanced to match crops nitrogen needs in which case phosphorus is applied excessively, which increases the accumulation of soil phosphorus [9]. Higher soil phosphorus is strongly linked to increase in the loading of dissolved forms of phosphorus, readily and fully available for algae in receiving waters [32]. This trade-off is important as we should be controlling the loading of both nutrients in most of the eutrophied surface water areas: Gulf of Mexico [50, 51], Chesapeake Bay [43] or the Baltic Sea [8]. It is particularly important for Lake Erie where algae growth is determined almost solely by phosphorus and where the loading of dissolved phosphorus forms has increased dramatically in the recent decade [18]. Regulating nutrient surpluses from manure applications is further complicated by two aspects. First, the regulatory grip on manure applications on animal farms' own crop production regions is much stronger than on surrounding crop farms' land. Animal farm can be made liable on its manure whereas the surrounding crop farms enter the world of manure on voluntary basis. Second, transboundary character of nutrient loading can create regional conflicts in regulatory emphasis of one nutrient over the other.

In this paper, we postulate a simple theoretical model featuring two most comprehensive features of manure management: physical coupling of nitrogen and phosphorus; and economic and regulatory division of animal and crop farms. We consider regulation which de facto focuses on only the nutrient relatively more scarce in manure with respect to crop's agronomic needs, and which binds only manure application on animal farm's own crop production area. With a sequence of propositions, we show that

i When faced with land constraints and discontinuities in benefits from manure utilization, increased environmental pressure does not necessarily lead to increase in manure utilization, even in social optimum
ii Single-nutrient approach may lead to increases in the surpluses of another via various channels
iii If constraining only one nutrient, nutrient management plans are not always beneficial for the environment. By extending the model we also show that

iv When nutrient surpluses also affect downstream regions, single-nutrient regulation may increase downstream externalities and even decrease the total welfare.

Freshwater ecosystems and coastal areas are often considered sensitive to phosphorus loading, and open sea areas to nitrogen [40]. Toxic mass blooms of blue green algae are a symptom of eutrophication which have received a lot of public attention. Their occurrence is linked to excessive phosphorus loading as their growth is not limited by nitrogen: blue-green algae is able to utilize atmospheric nitrogen [48]. Some consider phosphorus to be the most important long term driver of coastal water eutrophication as the atmospheric nitrogen fixation tends to gradually elevate the availability of nitrogen in the water ecosystem to the level where phosphorus becomes the limiting factor for algae growth [7]. In addition to the occasional mass blooms of blue-green algae, eutrophication has more persistent and equally economically severe symptoms such as permanent changes in fish stocks, increased turbidity of water or emergence of large anoxic sediment areas, so called dead zones.

Agriculture is the dominant source of nutrient loading to most U.S. surface waters [6]. It is estimated to contribute to 49 and 43% of nitrogen and phosphorus loading to the Chesapeake Bay and about 70% of both nutrients to the Gulf of Mexico [3, 55]. Sustainable manure management is one of the key issues in improving the nations water quality [22].

Large animal facilities tend to rely on imported feeds, i.e. imported nutrients [45]. Because high transportation costs erode its value as a fertilizer, manure generated in the facility is typically applied only up to a certain distance [17, 41]. Over-application of manure generates an immediate risk for nitrogen and phosphorus loading [46, 47]. It also gradually elevates the levels of phosphorus accumulated in the soil. Soil phosphorus is directly linked to loading of dissolved phosphorus, the fraction which has the strongest effect on eutrophication of phosphorus limited water ecosystems [10, 20, 32, 42]. Furthermore, a difference in N-P ratios of manure and crop's agronomic needs leads to even greater over-application of the nutrient relatively more abundant in manure. Applications are typically based on nitrogen content of manure and nitrogen need of the crops, leading to excessive phosphorus application even in fields with precise nitrogen usage. This further aggravates the phosphorus accumulation problem.

In the United States, regulation embodies the above division between farm types. Large animal facilities (CAFOs) are considered point sources which fall under the federal Clean Water Act whereas the crop farms as non-point sources are regulated by the states.[1] The difference in regulatory type and force is stark between the two farm types. Manure can be applied on animal farm's own fields (henceforth: on-farm) and on the fields of the surrounding crop farms (henceforth: off-farm). CAFOs on-farm applications must be planned and reported to match the needs of one of the crop nutrients. But when manure crosses the border of animal and crop farm, it by and large goes off the regulatory radar. Negative environmental effects, however, are not restricted to farm borders. Both on- and off-farm, the negative externalities are

[1] However, the specific rules and limits applied to CAFOs' manure management are state-specific.

ultimately caused by the nutrient balance: the difference of phosphorus (nitrogen) applied to the field and phosphorus (nitrogen) biologically uptaken by the crop.

Environmental regulation of animal agriculture is indeed focused on nutrient surpluses, though typically on either nitrogen or phosphorus. Conservation measures may lower both surpluses (e.g. having less animals or applying a given quantity of manure on a larger area) or they may affect the surplus of one nutrient only (e.g. altering the ratio of N and P in manure by feed choices, manure storage or application methods or switching to a crop with a different N-P ratio). The latter measures may increase or decrease the surplus of the other nutrient. The nutrient abatement measures may thus be substitutes or complements as determined by [28].

The earliest papers on animal waste regulation didn't differentiate between on-farm and off-farm application (or regulation), but included an outside disposal area. It was assumed to take care of manure with certain costs without environmental externalities [15, 27]. The assumption of safe disposal areas has not held in practice. The regional dairy management plan of San Jacinto, California, for instance, states that dairy farmers cannot know whether the hauling contractors actually apply the manure appropriately or whether they dump it illegally. The report also suggests that illegal dumping indeed takes place [38]. Huang et al. [16] assume that livestock farms lease the extra land needed for manure application, in which case they are also covered by on-farm regulation. The models by [17, 41] quantify the excess application of manure as a function of the number of animals and the distance from the facility. They do not allow for outside disposal areas and assume that regulation applies to on- and off-farm similarly.

The other alternative is to view on-farm and off-farm applications separately. Regulation applies to animal farms who may respond by altering their on-farm operations or by increasing off-farm export of manure. The on-farm responses might include changing the number of production animals, altering the nutrient composition of manure or manure handling and application practices, or changing the crop. These choices either alter the amount of nitrogen and phosphorus generated on-farm, the amount made available to crops or the amount uptaken by crops. Implicitly or explicitly, all papers, to our knowledge, assume in these cases that off-farm applications are done according to either agronomic needs of crops, as a perfect substitute for chemical fertilizers (see, for example, [13, 19] or [4]). Hence, the economic decision making for off-farm choices is not modeled explicitly.

We extend the literature by combining three elements into manure regulation models: setting apart and being explicit about on-and off-farm manure application choices and regulation; having the farms' choices simultaneously determine the phosphorus and nitrogen surpluses; and making regions' surface water quality partly dependent on other regions' choices. As seen above, each of these has been considered separately. There are, however, no models that combine all three in the context of animal manure and water quality regulation.

We develop a stylized model comprising two adjacent farms animal and crop farm and two pollutants phosphorus and nitrogen. The animal farm generates manure which it applies on its own fields (on-farm), and/or exports to be applied at the adjacent crop farm (off-farm) and/or dumps it. We solve for privately and socially

optimal manure application, animal numbers and crop choices. We then extend the model by acknowledging the transboundary character of water pollution. We include an independent downstream region which receives part of the residual nutrients. The damage caused by nitrogen and phosphorus loading is region-specific, i.e. the partial derivatives of the two regions' externality functions with respect to phosphorus and nitrogen are different. We examine two types of social optima: one focusing only on source region's externalities and the other taking into account both regions' externalities.

We show that the lack of available land on-farm may lead to costly regulation that has at its worst no environmental benefits; that focusing on nitrogen surpluses may even increase phosphorus loading; and regional orientation in regulation in worst cases may lead to an overall decline in social welfare. Furthermore, we show how regionally defined nutrient management plans transmit these effects into regional and global nutrient surpluses and welfare.

The rest of the paper is organized as follows. We first present the basic model and derive the necessary optimality conditions and their key implications. We then extend the model with the focus on regional and global welfare, regionally implemented instruments and their welfare implications. The third section concludes and discusses the policy implications.

2 The Model

Consider two representative farms: an animal farm which has its own feed production area, and a crop farm. Their border is denoted by the horizontal line at distance d in Fig. 1. Both farms' fields are assumed rectangular and one unit wide. Manure is generated at the facility located along the rear edge of the animal farm (along $d = 0$). Manure can be applied on-farm and exported and applied off-farm.

We assume that on the entire acreage where manure is applied, it is applied exactly according to crops' agronomic needs for the relatively more scarce nutrient (phosphorus or nitrogen).[2] With this assumption, hauling distance unambiguously determines the quantity of applied manure nutrients. We classify the excessive manure, i.e. the difference of manure generated by production animals and manure applied according to crops' needs, as dumped.[3] We thus assume that crop growth does not respond to nutrient application exceeding the agronomic need. The underlying crop

[2]All crop farm's fields are assumed to be suitable for manure application. Assuming a smaller fraction to be suitable would change the link between the application distance and acreage. This has trivial effects on the results and will thus not be further considered here.

[3]The counterpart of dumping in models by [17] or [41] is the excessive manure application which is monotonically decreasing with the distance from the animal facility, until becoming zero at the threshold distance. Their crop response functions are increasing and concave, ours is a linear response and plateau, simplified to a fixed yield and fixed need of nutrients. Externalities in both models are due to the sum of over application. Therefore, we do not need to define the exact location for dumping, as long as it's either on-farm or off-farm.

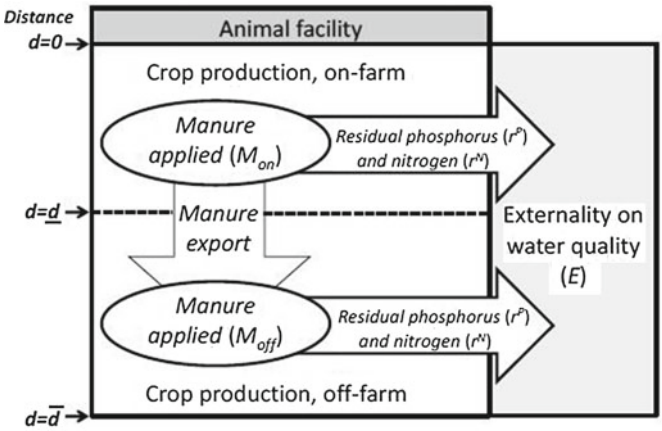


Fig. 1 Conceptual model of the agricultural region

response function is linear response and plateau (LRP). With fixed input and output prices, optimal fertilizer use leads to fixed, crop specific yields and associated nutrient requirements. Therefore, the application distance, crop choice and the nutrient content of manure together determine the quantity of manure applied.

Our model has two sources for residual nutrients from manure applications. First, the N-P ratios in manure and the N-P ratios of crops' agronomic needs differ. Therefore, applying manure according to crops' agronomic needs for the relatively more scarce nutrient generates a surplus for the other one. Second, dumping adds to residuals for both nutrients. To highlight the problems related to manure, we assume that chemical fertilizers are applied precisely according to crops' needs. Only manure may thus be the source of residual nutrients which are defined as:

Definition 1 Residual nitrogen $\left(r^N\right)$ and phosphorus $\left(r^P\right)$ are given by

$$r^N = \overbrace{\alpha q a}^{\text{Nitrogen generated}} - \overbrace{\gamma^k d_{on} - \gamma^j \left(d_{off} - \underline{d}\right)}^{\text{Nitrogen uptake}}$$

$$r^P = \overbrace{\beta q a}^{\text{Phosphorus generated}} - \overbrace{\delta^k d_{on} - \delta^j \left(d_{off} - \underline{d}\right)}^{\text{Phosphorus uptake}}$$

where α and β are the nitrogen and phosphorus concentrations of manure, q the amount of manure generated by one animal in one year, a the number of animals, γ and δ crop specific agronomic needs for nitrogen and phosphorus per acre, k and j the crop choices on- and off-farm, and d_{on} and d_{off} the hauling distances on- and

off-farm.[4] The application acreages are thus d_{on} and $d_{off} - \underline{d}$. Our model allows off-farm manure application even if there were no or only some on-farm application. That is, it is possible that $d_{off} > \underline{d}$ even if $d_{on} = 0$.

Nutrient residuals impair water quality. The damages may take multiple forms: elevated nitrate concentrations in ground water, mass blue-green algal blooms in surface waters, etc. We express the externalities with a general damage function, which has nitrogen and phosphorus residuals as arguments: $E\left(r^P, r^N\right)$. We assume no cross-effects in damage: $\frac{\partial E}{\partial r^i} > 0$, $\frac{\partial^2 E}{\left(\partial r^i\right)^2} > 0$, $\frac{\partial^2 E}{\partial r^P \partial r^N} = 0$.

Definition 2 The quantities of manue applied on-farm (M_{on}) and off-farm $\left(M_{off}\right)$ are given by:

$$M_{on} = max\left\{\frac{d_{on}\gamma^k}{\alpha}, \frac{d_{on}\delta^k}{\beta}\right\}$$

$$M_{off} = max\left\{\frac{\left(d_{off} - \underline{d}\right)\gamma^j}{\alpha}, \frac{\left(d_{off} - \underline{d}\right)\delta^j}{\beta}\right\}$$

The quantity of manure dumped on-farm (x_{on}) and off-farm $\left(x_{off}\right)$ are given by:

$$x_{on} + x_{off} = qa - M_{on} - M_{off}$$

Dumping on-farm (x_{on}) and off-farm $\left(x_{off}\right)$ have identical environmental effects. Because of transportation costs and the LRP crop response function, privately optimal dumping takes place onfarm with zero costs (technically in our model, on the shortest distance possible $d = 0$). We include the possibility for off-farm dumping because as we will see, it may be an optimal response to regulation.

We consider the optimal choices of (1) the animal farm, (2) the crop farm, (3) the social planner. The social planner maximizes the sum of profits from farming net of environmental damages, the crop and animal farms maximize profits.

Optimization Problem of the Crop Farm

The crop farm maximizes profits by choosing the crop and the amount of manure it imports as a substitute to chemical fertilizers (the total quantity of nutrients needed is fixed by the crop choice). It takes the price of manure as given.

$$Max_{j, d_{off}} \pi_{off} = \overbrace{y^j p^j \left(\bar{d} - \underline{d}\right)}^{\text{Sales revenues}} - \overbrace{\left(\gamma^j p^N + \delta^j p^P + g\right)\left(\bar{d} - d_{off}\right)}^{\text{Fertilization costs}} - \overbrace{p^M M_{off}}^{\text{Manure costs}} \quad (1)$$
$$s.t. d_{off} \geq \underline{d}$$

[4] Agronomic nutrient needs may differ from nutrient uptake of crops. Soybeans, for instance, can bind most of the nitrogen it needs from atmospheric nitrogen. We define residual nutrients as differences between actual applications and application requirements.

The crop is denoted by j and the amount of manure imported (or equally: the distance of manure application) by d_{off}.[5] The per acre crop yield is y^j, its net price (including variable costs other than fertilization costs) , and the total field acreage $\left(\bar{d} - \underline{d}\right)$. Costs from chemical fertilizers are given by $\left(\gamma^j p^N + \delta^j p^P + g\right)$ $\left(\bar{d} - d_{off}\right)$ where p^N and p^P prices of nitrogen and phosphorus, and g is the per-acre cost of application. The more manure applied, i.e. the longer the hauling distance d_{off}, the higher the savings from avoided chemical fertilization costs. With manure applied on the entire crop land (i.e. $d_{off} = \bar{d}$), chemical fertilization costs would be zero.

Crop Farm's Optimal Manure Import and Crop Choice

Writing a Lagrangian and taking the first-order conditions yields

$$\begin{aligned} \left(y^j p^N + \delta^j p^P + g\right) &= p^M \, max\left\{\tfrac{\gamma^j}{\alpha}, \tfrac{\delta^j}{\beta}\right\} + \lambda_c \\ \lambda \geq 0, \left(d_{off} - \underline{d}\right) \geq 0, \lambda_c \left(d_{off} - \underline{d}\right) &= 0 \end{aligned} \tag{2}$$

For a given crop, all terms in (2) are exogenous for the crop farmer. For positive manure import quantities ($d_{off} > \underline{d}, \lambda_c = 0$), a price the crop farmer is willing to pay is

$$p^M = \frac{\left(\gamma^j p^N + \delta^j p^P + g\right)}{max\left\{\tfrac{\gamma^j}{\alpha}, \tfrac{\delta^j}{\beta}\right\}} \tag{3}$$

Given fertilizer prices, crop choice, and manure nutrient concentration, the price (3) is constant. That is, the crop farmer is always willing to use manure for the entire suitable crop land or not at all. The eventual amount of imported manure depends on animal farm's willingness to sell manure, which is driven by the hauling and application costs.

The price (3) is increasing in nutrient concentration of the relatively scarce nutrient and insensitive toward the other nutrient. However, the price is affected by prices of both nutrients as chemical fertilizers. The numerator in (3) gives costs in dollars per acre of chemical fertilization, which is influenced by both nutrients and the application costs. The denominator gives the amount of manure one needs to cover the nutrient requirements of an acre of land. Hence, the unit of (3) is dollars per unit (e.g. gallon) of manure. There is a different threshold price for each crop.

Optimization Problem of the Animal Farm

The animal farm chooses the number of animals, manure application on own land, manure export and the cultivated crop to maximize profits:

[5]Hauling distance is a common metric for crop and livestock farmer. For tractability, we denote hauling distances with subscripts. That is, if the crop farm's hauling distance is equal to the boundary of the animal and crop farm ($d_{off} = \underline{d}$), it does not import manure.

$$\underset{a,d_i,x_{off},k}{Max}\ \pi_{on} = \overbrace{p^a a + p^M M_{off}}^{\text{Sales revenues}} - \overbrace{f(a)}^{\text{Production costs}} - \overbrace{\left(\gamma^k p^N + \delta^k p^P + g\right)\left(\underline{d} - d_{on}\right)}^{\text{Fertilization costs}} \quad (4)$$
$$- \overbrace{p^k\left(\xi^k a - y^k \underline{d}\right)}^{\text{Feed costs}} - \overbrace{h\left(M_{on}, d_{on}\right) - h\left(M_{off}, d_{off}\right)}^{\text{Hauling costs}} - \overbrace{c(x_{off})}^{\text{Dumping costs}}$$
$$s.t.\ \left(qa - M_{on} - M_{off}\right) \geq 0;\ \left(\underline{d} - d_{on}\right) \geq 0;\ \left(d_{off} - \underline{d}\right) \geq 0;\ \left(\bar{d} - d_{off}\right) \geq 0$$

Primarily, the revenues are obtained from production animals (a). Potentially, there may also be revenues from selling manure and feed grown in animal facility's own fields (given production exceeds the needs of production animals). Substituting fertilizers with manure also creates savings. Life-cycle revenues from a single animal are given by p^a. Depending on the type of production animal, these may comprise average per-unit revenues from selling milk, meat, eggs etc. Manure $\left(M_{off}\right)$is sold with a price p^M, which is determined in (3). Fertilizer savings from applying manure on d_{on} are given by $d_{on}\left(\gamma^k p^N + \delta^k p^P + g\right)$.

Production costs (f) comprise of annualized investment costs and operation costs excluding feed costs. The cost function satisfies $f'(a) > 0$ and $f''(a) > 0$.[6] Feeding (a) animals requires $\left(\xi^k a\right)$ units of forage. The forage requirement depends on the animal and on the crop. Own crop production $\left(y^k \underline{d}\right)$ may be higher or lower than this. The needed (excess) units of feed will be bought (sold) at price p^k. This, of course, is a substantial simplification of the complex problem of how to feed the animals, how much and what to import and how much and what to grow in farm's own fields. However, this is not a crucial part of the model for the insights we derive from it. For us it is important that the crop is actively chosen and its N-P requirements together with manure application and its N-P concentration make up the nutrient surpluses.

Hauling and application costs (h) depend on the distance and the quantity of manure hauled. In our model, the latter is determined by distance and crop choice. Following conventional assumptions (see, for instance [13]) we assume that the costs of hauling a unit are increasing in distance and thus total hauling costs are increasing and convex.[7] The function is identical for on- and off-farm hauling, but off-farm hauling starts from a distance $\underline{d}$.

The costs of dumping (c) on own land are assumed to be zero and, on the crop farm, $c\left(x_{off}, \underline{d}\right) \equiv x_{off}\frac{\partial h(\underline{d})}{\partial d}$. Trivially, $x_{off} = 0$ without environmental regulation. The first constraint in (4) limits the total amount of manure applied to manure generated $\left(qa - M_{on} - M_{off}\right)$, associated with a shadow price λ_1 in the constrained maximization problem; the availability of on-farm land $\left(\underline{d} - d_{on}\right)$, shadow price λ_2; technical constraint for off-farm hauling distance to be at least $\underline{d}$, shadow price λ_3 and the availability of off-farm land for manure application $\left(\bar{d} - d_{off}\right)$, shadow price λ_4.

[6] Think of f as a simplification from a concave-convex cost function. The sufficient (second-order) conditions tell that the relevant part of the curve must have a positive second derivative.

[7] We could also assume linear or even concave hauling costs. This would make the optima characterized by some binding constraints. Qualitatively, the results would remain unchanged.

Optimal Animal Numbers, Crop Choice, Manure Utilization and Export

The first-order conditions (excluding, for brevity, the standard non-negativity constraints) for the continuous choice variables are

$$p^a + \lambda_1 q = f'(a) + p^k \xi^k \tag{5}$$

$$\frac{\partial h}{\partial d_{on}} = \left(\gamma^k p^N + \delta^k p^P + g\right) - \lambda_1 max\left\{\frac{\gamma^k}{\alpha}, \frac{\delta^k}{\beta}\right\} - \lambda_2 \tag{6}$$

$$\frac{\partial h}{\partial d_{off}} = p^M max\left\{\frac{\gamma^j}{\alpha}, \frac{\delta^j}{\beta}\right\} - \lambda_1 max\left\{\frac{\gamma^k}{\alpha}, \frac{\delta^k}{\beta}\right\} + \lambda_3 - \lambda_4 \tag{7}$$

$$\lambda_1 \geq 0, \left(qa - M_{on} - M_{off}\right) \geq 0, \lambda_1 \left(qa - M_{on} - M_{off}\right) = 0$$

$$\lambda_2 \geq 0, \left(\underline{d} - d_{on}\right) \geq 0, \lambda_2 \left(\underline{d} - d_{on}\right) = 0$$

The optimal number of animals (5) balances the marginal benefits and costs of having one more animal. The marginal benefits consist of sales revenues (p^a) and the shadow value of manure ($\lambda_1 q$). If manure is scarce ($\lambda_1 > 0$) it would be applied more if available. If manure is excessive, ($\lambda_1 = 0$), benefits accrue from sales revenues only. The marginal costs consist of marginal production costs $\left(f'(a)\right)$ plus feed costs for one animal $\left(p^k \xi^k\right)$), i.e. a linear and decreasing function.[8]

The optimal hauling distance on-farm (6) balances the marginal savings in mineral fertilizers $\left(\gamma^k p^N + \delta^k p^P + g\right)$ and the marginal costs of hauling. If manure quantity and area constraints are not binding ($\lambda_1 = \lambda_2 = 0$), marginal costs equal marginal savings in chemical fertilizer use. If manure is applied on the entire farmland controlled by the livestock farm, and more would be applied were the area larger, the area constraint is binding: $\lambda_2 = \left(\gamma^k p^N + \delta^k p^P + g\right) - \frac{\partial h}{\partial d_{on}} > 0$.

Conditions for optimal hauling distance off-farm (7) are similar except that the marginal benefit is the price received from the crop farmer (3). If crop choices on- and off-farm are identical, the shadow value of the land constraint is the negative of the shadow value for the animal farm's land ($\lambda_3 = -\lambda_2$).[9] If the land constraint on crop farm is binding (λ_4), the entire agricultural region is not sufficiently large to absorb generated manure nutrients. The opposite is not true however. Even though the region would not be enough to absorb all generated nutrients, the land constraint might not be binding.

Combinations of binding and non-binding manure and land constraints on- and off-farm generate eight different cases. The optimality conditions simplify differently

[8]Note that the farmer marginally loses $\left(p^k \xi^k\right)$ whether the farmer is a net importer or exporter of feed. If the farmer produces more than the production animals need, increasing the number of animals reduces the sales revenues. If the farmer has to buy the additional feed needed, the input costs increase by the same amount.

[9]This would be different if the animal farm did not retrieve the entire surplus from the crop farm in selling manure: its gains from both land applications are identical.

for each of them. We narrow our focus on the most policy relevant cases: The manure is excessive ($\lambda_1 = 0$), nitrogen is the relatively scarce nutrient, and the animal farm utilizes at least some of the manure and may or may not export it to the crop farm.

Optimization Problem of the Social Planner

The social planner maximizes private profits from operations on- (π_{on}) and off-farm $\left(\pi_{off}\right)$ net of externalities (E).

$$\underset{a,d_i,x_i,j,k}{Max}\ \pi_{on}(a,k,d_i,x_i) + \pi_{off}\left(j,d_{off}\right) - E\left(r^N,r^P\right) \tag{8}$$

Optimal Animal Numbers, Crop Choices and Manure Utilization

denoting the shadow values with λ_i^E, social planner's first order optimality conditions are:

$$p^a = f'(a) + p^k\xi^k + \frac{\partial r^N}{\partial a}\frac{\partial E}{\partial r^N} + \frac{\partial r^P}{\partial a}\frac{\partial E}{\partial r^P} \tag{9}$$

$$\frac{\partial h}{\partial d_{on}} = \left(\gamma^k p^N + \delta^k p^P + g\right) - \lambda_2^E - \frac{\partial r^N}{\partial d_{on}}\frac{\partial E}{\partial r^N} - \frac{\partial r^P}{\partial d_{on}}\frac{\partial E}{\partial r^P} \tag{10}$$

$$\frac{\partial h}{\partial d_{off}} = \left(\gamma^j p^N + \delta^j p^P + g\right) - \lambda_3^E - \lambda_4^E - \frac{\partial r^N}{\partial d_{off}}\frac{\partial E}{\partial r^N} - \frac{\partial r^P}{\partial d_{off}}\frac{\partial E}{\partial r^P} \tag{11}$$

$$x_{on} \geq 0, x_{off} = 0 \tag{12}$$

Given that the social planner chooses the same crops as the private farmers, the optimal number of animals decreases: the two last terms in (9) are positive, $f'(a) > 0$ and other terms are constant. Changes in hauling distances, (10) and (11), are slightly different as they are influenced by the border of crop and animal farms. This is one of our inputs to traditional animal waste models:

Proposition 1 *If the marginal social benefits from manure utilization at the crop and animal farm's border are below marginal hauling and application costs; and if the animal farm fully utilizes its land application area in private optimum, social planner's solution does not increase manure utilization. That is, if* $\lambda_3^E > 0$ *and* $\lambda_2 > 0$ *then* $d_{on} = \underline{d}$ *and* $M_{off} = 0$ *for both private and social optima.*

Figure 2 illustrates the proposition. The horizontal axis denotes the hauling distance, the marginal gains and costs from manure application are on the vertical axis. The dotted vertical line at $d_{on} = \underline{d}$ denotes the border of the animal and crop farm. The increasing curve denotes the marginal hauling costs. The lower horizontal line on the left of the dotted vertical line denotes the private marginal benefits from manure application on-farm $\left(\gamma^k p^N + \delta^k p^P + g\right)$. More manure would be spread on-farm if more land was available. Hence, the shadow price for land on-farm ($\lambda_2 > 0$)

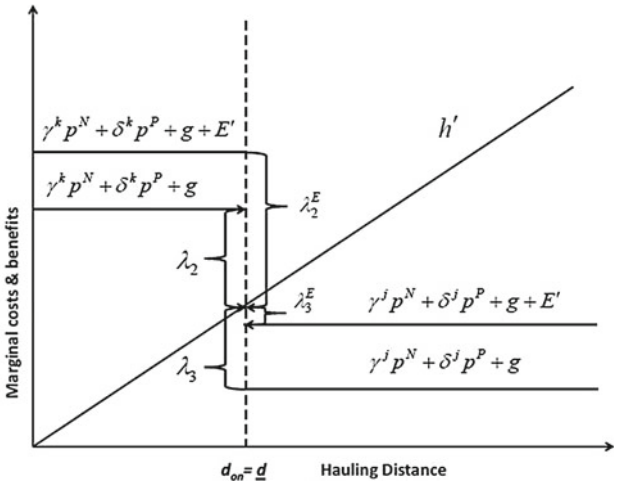


Fig. 2 Optimal hauling distances on- and off-farm

is positive. Introducing externalities (the upper horizontal line) puts more pressure on utilizing manure on-farm, but as all land is already used for manure application, the only effect is for the shadow price for on-farm land to increase $\left(\lambda_2^E > \lambda_2\right)$. Off-farm, no manure is applied in the private optimum, i.e. $\left(\gamma^j p^N + \delta^j p^P + g\right) < h'$ and $\lambda_3 > 0$. The socially optimal shadow price of crop land $\left(\lambda_3^E\right)$ decreases close to zero but the actual utilization of manure does not change as the hauling distances remain unchanged. Externalities thus increase the need for higher utilization of manure nutrients, expressed in shadow prices, but the availability of on-farm land may result in unchanged manure utilization levels. Of course, if the crop choices on- and off-farm were the same, the marginal social benefits of manure application would be identical on- and off-farm. Therefore, an increase in marginal damage would always lead to longer hauling distances.

In our parsimonious model, there is no reason for the animal and crop farm to choose their crops differently. In practice, however, this is often the case. Furthermore, there are other features that would be associated with differences in marginal benefits from manure application on- and off-farm even if the crop choices were identical. There might be discontinuities in hauling distances or crop farmers might be unwilling to apply manure on (some of) their fields.

The externalities from dumping manure either on-farm or off-farm are identical. Since the costs of dumping on-farm are assumed to be zero, the social planner chooses trivially $x_{off}^* \geq 0$. The amount dumped is defined by the other optimal choices: $x_{on}^* = qa^* - M_{on}^* - M_{off}^* \geq 0$.

3 Instrument Analysis

By definition, any global sub-optimality of rural planner's policies will carry over with first-best instruments imposed on farmers. The situation is different with second-best instruments, of which we analyze Nutrient management plans (NMP) and a tax on nitrogen fertilizer. NMPs are central in regulating CAFOs in the United States. Similar, nutrient balance based approached are also widely used in other countries. Tax on nitrogen fertilization is an economic instrument that could be considered for nutrient regulation and it has been used in, for instance, Austria, Finland and Sweden [37]. We analyze these from the rural social planner's perspective and examine how the unintended increase in phosphorus residuals is affected by the instruments.

Nutrient Management Plans

NMPs may be based either on nitrogen or phosphorus standard. In both cases, the optimization problem for the animal farm changes. The manure application distance is simply set either at $d_{on}^{N} = \frac{\alpha q a}{\gamma^k}$ (nitrogen standard), at $d_{on}^{P} = \frac{\beta q a}{\delta^k}$ (phosphorus standard) or at $d_{on} = \underline{d}$ (not enough on-farm land). The optimal responses to NMPs as well as the unregulated optima are summarized in the following Proposition. Assuming that the animal farm does not respond to the NMP by switching crops, and denoting the optimal choices under nitrogen standard with the superscript NMPN, under phosphorus standard with NMPP and under private optimum with $\left(^{\wedge}\right)$ we obtain:

Proposition 2 *Under NMP, with nitrogen as the relatively scarce nutrient, it holds for any given crop that*

i $a^{NMPP} \leq a^{NMPN} < \hat{a}$
ii $d_{off}^{NMPP} = d_{off}^{NMPN} = \hat{d}_{off}$
iii $\hat{d}_{on} < d_{on}^{NMPN} \leq d_{on}^{NMPP}$
iv $x_{off}^{NMPN} < x_{off}^{NMPP} < \hat{x}_{on}$
v $\pi^{NMPP} < \pi^{NMPN} < \hat{\pi}$

Consider two cases for (i). If on-farm application area is fully utilized, the additional component to the right side of (5) $\left(q\frac{\partial h(\underline{d})}{\partial d}\right)$ is identical under both standards and the optimal number of animals decreases equally under both standards. If on-farm land is sufficient for manure application satisfying the NMP, the additional component is $\frac{\partial h(\underline{d})}{\partial d}\frac{q\alpha}{\gamma^k}$ under nitrogen standard and under phosphorus standard. As we consider the case where nitrogen is the relatively scarce nutrient $\left(\text{i.e.}\frac{\gamma^k}{\alpha} > \frac{\delta^k}{\beta}\right)$, phosphorus standard adds a larger component to the right side of (5) and the optimal number of animals is smaller. The intuition is clear: hauling costs are larger for each additional manure unit under phosphorus standard and its negative effect on profitability of animal husbandry is thus stronger.

The second point (ii) is almost trivial, yet powerful. Neither of the standards changes the economic decision making or the regulatory environment of the crop farm

and does therefore not increase off-farm utilization of manure. This is a key intuition of our model: Without regulatory grip of crop farms and with fixed manure handling technologies, on-farm nutrient standards do not increase utilization of manure outside animal farm's borders. However, if the animal farm responds to NMPs by investing in technologies decreasing hauling costs or promoting the demand off-farm, off-farm utilization does increase. It would be essential to find out how animal farms have actually been responding to constraints posed by NMP.

For (iii) consider two cases. If on-farm land is fully utilized, nutrient standard cannot change the on-farm hauling distance $\left(d_{on} = \underline{d} \text{ for both standards}\right)$. If on-farm land is only partly utilized, same reasoning as above holds: land-application of manure increases more under P-standard. This is a familiar outcome. For instance [2] show that under phosphorus standard the required land area for manure application is substantially larger than under N-standard.

There are differences between the two standards in how much is dumped off-farm as shown by (iv). It states that the amount of manure dumped off-farm under phosphorus standard is higher than under the nitrogen standard, both being lower than dumping under no regulation. This is an important result concerning the differences between the two standards. Recall that, even though the constraint of $x_{on} = 0$ still remains, the maximization problem changed slightly when moving to a phosphorus standard. It turns out that less manure will be applied on-farm and as off-farm hauling is unchanged, the number of animals is lowered identically by both standards, off-farm dumping must increase. The costs increased with fertilization costs and increased dumping to the farm border, the result referred to in (v) above.

There are two more features to be noted. First, NMPs incentivize switching to a higher nitrogen or phosphorus uptaking crop, depending on the standard. Therefore, insights of Proposition 2 apply also here. Second, even though the phosphorus standard will increase crop-specific dumping, we do not know its total environmental or welfare effects: The crop switch affects these as well as parametrizations of the relevant functions.

Proposition 2 contradicts with [19] who find that binding nutrient constraints reduce phosphorus loading significantly. The difference stems from the modeling assumptions: our model assumes that an animal farm follows either nitrogen or phosphorus standard as is the case with actual CAFO regulation. Furthermore, our model assumes that the animal farm may comply by exporting all surplus manure off-farm, i.e. export and apply more than is paid for. In the partial equilibrium model by [19], the crop farm accepts only the amount of manure required by crops. Hence, the difference in how binding the regulation is on-farm and off-farm changes the results drastically.

The substitution between phosphorus and nitrogen surpluses is similar to that of air and water pollution in [1] In their model, applying manure without incorporating it to soil reduces nitrogen surpluses contributing to ground water pollution but increases air pollution, both topical problems in California. Effectively, this is the same effect as that brought by switching the crops in our model in terms of phosphorus and nitrogen surpluses. Also [4] consider cross-effects of air and ground-water pollution but not those of phosphorus and nitrogen.

Nitrogen Tax

We analyze a tax on mineral nitrogen fertilizer in the simplest possible setting: the animal farm applies manure according to nitrogen needs, it dumps some of the manure it generates and it does not export any manure before or after the tax. Furthermore, to gain analytically insightful results, we parametrize the hauling costs as $h = M\phi d_i$ where ϕ is some parameter. As nitrogen is the relatively scarce nutrient, hauling costs are given by $h = M\phi d_i = \frac{\phi d_i^2 \gamma^l}{\alpha}$, i.e. increasing and convex in distance and decreasing in the concentration of the relatively scarce nutrient.

Because the model contains continuous and discrete variables, we analyze the effects of a tax in two steps. First, we examine how it would change the animal farmer's optimal choices regarding the number of animals and hauling distance. Then, we examine what kind of incentives it creates for crop choice.

A nitrogen tax increases the price of nitrogen fertilizers. Comparative statics reads:

$$\begin{bmatrix} f'' & 0 \\ 0 & \frac{-2\phi\gamma^k}{\alpha} \end{bmatrix} \times \begin{bmatrix} \frac{da}{dp} \\ \frac{dd_{on}}{dp} \end{bmatrix} = \begin{bmatrix} 0 \\ -\gamma \end{bmatrix}$$

Yielding $\frac{da}{dp} = 0$ and $\frac{dd_{on}}{dp} = \frac{\alpha}{2\phi} > 0$. An increase in the price of nitrogen increases the hauling distance at the rate of the ratio of nitrogen concentration in manure and marginal hauling costs. Hence, it decreases the residuals of both nutrients, given that there are no changes in crop choice. What kind of incentives does a tax on nitrogen create for crop choice? A tax increases the per-acre costs of chemical fertilization and, therefore, makes it profitable to haul and apply manure on a larger area. The higher the crop requirement for nitrogen, the higher the marginal effect of fertilizer price increase on profits. That is, increasing nitrogen fertilizer prices creates incentives to change the crops to those requiring less nitrogen. The rural social planner, trying to lower the nitrogen residual, does not want to see the farmer to respond to τ by switching to a crop that requires less nitrogen. This would increase the nitrogen residuals from the given manure application (the effect on phosphorus residuals depends on the phosphorus uptake of the new crop). The rural social planner is thus willing to set a tax (τ) on a range $0 \leq \tau < \tau_k$, where τ_k is given by the equality of any alternative crop choice (s) such that

$$\begin{aligned} \pi_\tau^k - \pi^s = 0 \Leftrightarrow \left(\gamma^k \left(p^N + \tau_k\right) + \delta^k p^P + g\right)\left(\underline{d} - d_{on}\right) + p^k \left(\xi^k a - y^k \underline{d}\right) + h\left(M_{on}^k, d_{on}^k\right) \\ = \left(\gamma^s p^N + \delta^s p^P + g\right)\left(\underline{d} - d_{on}\right) + p^s \left(\xi^s a - y^s \underline{d}\right) + h\left(M_{on}^s, d_{on}^s\right) \end{aligned}$$

That is, the tax is bound from above at the level where the farmer is indifferent between switching the crops. For a given crop, a tax does not incentivize increasing (or decreasing) the number of animals or encourage a transition to a crop with higher nitrogen uptake crop. The effectiveness of the nitrogen tax is an empirical question.

Reference [41] suggest that an increase in fertilizer price increases both manure application and the number of production animals. The difference with our model stems from the crop response specification. They assume that marginal impact of

increase in nutrients is always positive. In our model with fixed crop yields, a price increase affects only the hauling distance (when manure is excessive). Contrasting both these models, [49] find that a 56% increase in nitrogen price does not affect manure utilization in crop production.

4 Transboundary Pollution Case of Two Regions

Nutrient pollution to surface waters is transboundary: Phosphorus and nitrogen emitted upstream may impair water quality at the source areas or at any location along the river or at the sea. The environmental damage at affected locations depends on the aquatic ecosystems, and on the distance between the source and the receptor area. If regions have independent jurisdictions, downstream regions tend to suffer from externalities excessively as upstream regions free ride.

We want to focus on a situation where the upstream region does internalize the externalities in its own region but following from the interconnectedness of phosphorus and nitrogen surpluses in animal agriculture and differences in environmental characteristics this has unintended welfare implications. Let us extend the model to include a region located downstream from the agricultural region. The downstream region has no polluting activities of its own but it receives a fraction $0 \leq \omega_N \leq 1$ of agricultural region's nitrogen surplus $\left(r^N\right)$ and $0 \leq \omega_P \leq 1$ of its phosphorus surplus $\left(r^P\right)$.[10] There, they generate externalities according to a region-specific function $E^D\left(\omega_N r^N, \omega_P r^P\right)$. We consider three types of optima: Private optimum that does not consider externalities, agricultural region's optimum (henceforth rural optimum), determined by the rural social planner solution to (8) that solves the inboundary externality but ignores the transboundary pollution and the global planner's optimum that maximizes the agricultural region's profits net of environmental damages in both regions[11]:

$$\underset{a,d_i,x_i,j,k}{Max}\ \pi_{on}(a,k,d_i,x_i)+\pi_{off}\left(j,d_{off}\right)-E\left(r^N,r^P\right)-E^D\left(\omega_N r^N,\omega_P r^P\right) \tag{13}$$

The globally optimal number of animals is given by:

$$p^a=f'(a)+p^k\xi^k+\frac{\partial r^N}{\partial a}\left(\frac{\partial E}{\partial r^N}+\omega_N\frac{\partial E^D}{\partial r^N}\right)+\frac{\partial r^P}{\partial a}\left(\frac{\partial E}{\partial r^P}+\omega_P\frac{\partial E^D}{\partial r^P}\right) \tag{14}$$

[10] For simplicity, we assume the fractions exogenous. In reality, they reflect the amount of externalities in the agricultural region: algae growth, for instance, reduces the amount of nutrient residuals that are eventually carried over to recreational region's surface waters. Intensive algae growth in the agricultural region would lower both ω_N and ω_P.

[11] As there are no polluting activities in downstream region, its optimum would be trivially: $r^N = r^P = 0$.

Conditions for globally optimal hauling distances on- and off-farm (not presented here) include the same partial derivatives of downstream region's externalities, weighted with the carry over fractions. If the environment in the downstream region is sensitive towards a nutrient of which some fraction is carried downstream, globally optimal solution differs from the rural planner's optimum.

Proposition 1 and the ensuing discussion hold for the global planner. Looking at Fig. 2, global planner's solution, given $\omega_i \frac{\partial E^D}{\partial r^i} > 0$ would shift the horizontal line associated with crop j higher.

There would, however, still be cases where discontinuities in marginal benefits from manure applications off-farm would cause the consideration of externalities in no improvements in manure utilization.

Let us examine how rural planner's intervention changes the amounts of residual nutrients and environmental damages in both regions. Let us first make the following notational definitions:

Definition 3 The differences in nitrogen and phosphorus residuals associated with rural planner's and private farmers' optima are denoted by $\Delta r^N = r^{N^*} - \hat{r}^N$ and $\Delta r^P = r^{P^*} - \hat{r}^P$, respectively. The ensuing difference in environmental damage downstream is denoted by ΔE^D.

Rural planner's regulation always decreases the externalities in the agricultural region, and hence the residual of at least the nutrient towards which its environment is more sensitive to. The following two propositions establish the conflict between the agricultural and downstream region.

Proposition 3 *Assume that rural planner's intervention reduces nitrogen residuals. This may result in an increase in phosphorus residuals in which case* $\Delta r^N \Delta r^P < 0$. *Then, for each* Δr^N, *there is some* $\Delta \tilde{r}^P$ *and some damage function parametrization for which* $\Delta E^D > 0$.

Proposition 3 states that, if reductions in nitrogen residuals are associated with increases in phosphorus residuals carried over to the recreational region, the externalities in the recreational region may increase as a result of rural planner's intervention. If there is a trade off in nitrogen and phosphorus residuals, and if the sensitivity of the environment is different in the two regions, there is some threshold after which rural regulation makes the downstream region worse off.

What are the conditions for $\Delta r^N \Delta r^P < 0$? First, the crop choices between the rural planner and the private farmers must be different. If they are identical, the residuals move in same directions. The rural planner always generates less (here) nitrogen residuals than the private farmers. If the per-acre crop uptakes are unchanged, the phosphorus residual has to decrease too.

Second, the differences in nitrogen and phosphorus uptakes of the crops must be high enough to offset the potentially longer hauling distances and the potentially decreasing number of production animals. It is thus not the absolute phosphorus uptake that matters. If the rural social planner switches to a crop with a significantly higher nitrogen uptake and with a slightly higher phosphorus uptake, the per-acre

phosphorus residual increases in the manure application area. This increases the phosphorus residual in both regions.

Whether residuals moving in opposite directions actually increase environmental damage downstream depends on the ecological characteristics of the surface waters. Phosphorus may be the only determinant of eutrophication or it might have no effect on algae growth or anything between the extremes. The more important it is relative to nitrogen, the more likely it is that the rural social planner's policies will deteriorate the environmental quality downstream. Also the ratio of carryover rates for phosphorus and nitrogen affects the effect: the higher the ratio $\frac{\omega^P}{\omega^N}$, the stronger the negative effect downstream.

The absolute levels of the coefficients are not important as long as they are nonzero. The same does not hold for the following proposition:

Definition 4 Denote by $\Delta\pi = \hat{\pi} - \pi^*$ the difference in profits associated with the rural planner's and private farmers' optima and by the associated difference in environmental damage in the agricultural region.

Proposition 4 *If $\Delta r^N \Delta r^P < 0$ and $\Delta E^D > 0$ there is some threshold increase in welfare in the agricultural region, $\Delta\tilde{W} = \Delta\pi - \Delta E$, that has to be achieved for rural social planner's intervention to increase global welfare.*

Proposition 4 compares the cases of no intervention and rural intervention from the global planner's perspective. If the rural planner's solution increases one of the nutrient surpluses while decreasing the other $\left(\Delta r^N \Delta r^P < 0\right)$; and if the environmental sensitivity of the downstream region is such that the environmental quality decreases as a consequence $\left(\Delta E^D > 0\right)$, rural planner's optimal solution may decrease global welfare. In this case, the no intervention case outperforms the rural intervention, from the global planner's perspective. Note that the rural planner always increases welfare and environmental quality in the agricultural region. But if it simultaneously increases environmental damages downstream, the welfare improvement in agricultural region must offset this effect. As the coefficients ω^P and ω^N become smaller, a decrease in global welfare becomes less likely. For the individual elements of welfare changes, discussion in Propositions 2 and 3 apply.

The results established in Propositions (2)–(4) bear similarities with [57]. Their theoretical model considers a single product with multiple externalities during its life cycle. The main outcome of their model is the need to consider all externalities from a single product simultaneously. The results are also related to the game theoretical model of [24] who illustrate a conflict between agricultural and ground water managers, drawing from their differing definitions of externalities. And even though we have not analyzed policy instruments so far, the results also hint towards the Tinbergen rule [54] which states that economic policy must include as many instruments as targets. After all, the ultimate reason for the regional conflict is the different role of nutrients in the environmental damage in the two regions, which in the extreme case is the complete lack of one of the externalities in the rural planner's optimization problem.

5 Discussion and Policy Implications

We developed a two-agent, two-pollutant, two-region animal waste management model, recognizing the difference between manure application on animal vs. crop farms, and considering simultaneously phosphorus and nitrogen surpluses and identifying possibility of conflicts between farming and non-farming regions with shared body of water. We assessed the implication of regulation using the popular nutrient management plan in addition to financial incentives. We showed that governance design crucially affects the outcomes of environmental regulation.

There are two insights provided by the on-farm off-farm division. Firstly, increasing pressure to reduce nutrient surpluses does not always lead to more efficient manure utilization. If the animal farm's fields are already fully utilized, introducing (or increasing) the environmental damage associated with nutrient surpluses does not necessarily lead to increased hauling distances.

The second insight is related to Nutrient Management Plans (NMP) as a regulatory instrument. NMPs provide binding nutrient surplus constraints to on-farm applications. The marginal incentives for manure utilization off-farm, however, do not change. NMPs do foster record keeping and information guidance for crop farmers utilizing manure (see e.g., [31] and Pennsylvania Code §83.343) The effectiveness of education or information as a regulatory instrument, however, is questionable (see, e.g., [36]). At its worst, NMPs might induce crop production areas to be used to get rid of excessive manure at application rates higher than agronomic recommendations. Hence, excessive manure applications might be simply shifted from one region to another, with minor benefits to environment but with increased hauling costs.

Our analysis indicates the existence of a trade off in nitrogen and phosphorus surpluses. This is relevant in the U.S. where about 90% of hogs and pigs and about 66% of milk cows are in operations classifiable as CAFOs [56]. [12] Any regulation that affects CAFOs' generation and application of manure has substantial consequences on nutrient loading on nation's surface and ground waters.

Our analysis has potential implications for regulating transboundary pollution to the Gulf of Mexico. The Gulf suffers from a persistent hypoxia area caused by eutrophication. During the last decades, its size has fluctuated around 15,000 square kilometers [52]. It causes huge direct losses to fisheries but is also contributes to the eutrophication itself by disallowing phosphorus to be trapped in bottom sediments.

Nutrients to the gulf are brought by the Mississippi, and Atchafalaya Rivers, collecting waters from over thirty states. Previously, nitrogen was considered the most important cause of eutrophication in the Gulf of Mexico (see, e.g., [34]. In the first Action Plan of the Mississippi River/Gulf of Mexico Watershed Nutrient Task force [50] it was explicitly stated that the excessive algal growth was primarily driven by nitrogen. Even though phosphorus and local water quality concerns were mentioned, the two priorities chosen to combat dead zones were linked to nitrogen

[12] CAFO definitions vary by state and do not match perfectly the classes of Agricultural Census data; 90% of hogs and pigs are on farms that have more than 2000 heads and 66% of milk cows are on farms with more than 200 heads.

loading. This, in turn, might have influenced the nutrient management plans the states in the basin adopted. Our results suggests that if NMP are based on a nitrogen standard, phosphorus loading from large animal facilities may increase for a variety of reasons.

Current understanding suggests that both nitrogen and phosphorus have important roles in driving eutrophication in the gulf [33, 39]. This is also recognized in the action plan: the latest version emphasizes both phosphorus and nitrogen abatement [51]. The 2013 progress report shows that the five year average nitrogen loading to the Northern Gulf of Mexico has remained at its 1997 level, whereas phosphorus loading has increased by about 30% [52]. Obviously, an increase of this magnitude is bound to have negative effects on the water quality of the Gulf of Mexico. We know that during this period, point-sources have further curtailed their pollution, the sales of mineral phosphorus fertilizers have not systematically increased and that the agglomeration and intensification of animal operations has continued.[13] It is plausible that manure management practices as described in the current paper could partly explain the increase. To empirically verify this, we would need an econometric analysis of crop prices and choices, fertilizer prices, manure standards followed, amounts of imported and exported manure, etc. A simple first step would be book keeping: how many of the CAFOs in the basin are following nitrogen and phosphorus standards and how has this changes over time as farm sizes have increased?

The third feature of our model was the regionally independently regulated, transboundary pollution. In the United States, the Environmental Protection Agency guides federal-level policies while the states have primacy in implementation and enforcement of regulations, such as the CleanWater Act. [44] proposes that this decentralization has, to some extent, lead to free riding by states. We show that regulation focusing on a single region and a single nutrient may lead to similar results as free riding, even in first-best optimum.

Transboundary pollution is often analyzed under the framework of environmental federalism. It analyzes benefits and drawbacks of local versus federal regulation. The basic result suggests that regions with independent environmental regulation tend to be driven toward overly lax environmental policies the race to the bottom hypothesis (see, for instance [21, 30]. Distortion from efficient levels of environmental protection are typically a result of strategic reasoning (local level regulation) or the lack of environmental precision and other informational shortcomings (federal level regulation). By introducing two tightly linked pollutants, our model generates similar outcomes under full information and without strategic play between the regions. So far, all theoretical frameworks on transboundary pollution and environmental federalism, starting from [5, 53] as well as later developments, such as [12, 23] focus on a single pollutant. This is hardly suitable for water quality issues driven by one macro nutrient in one place and by a combination of both in the other.

There are obvious extensions to our analysis. We could explicitly assume only an endogenous fraction of cropland suitable for manure application. There are technical and crop-specific reasons for suitability but also reluctance of crop farmers to apply

[13] See, e.g., http://www2.epa.gov/nutrient-policy-data/commercial-fertilizer-purchasedtable2.

manure on their crops. An often cited reason for farmers' unwillingness to accept manure is their uncertainty regarding the nutrient concentration of manure and the plant availability of these nutrients. Crop farmers' willingness to accept manure is often found to be crucial for livestock farmers' compliance costs [19, 35]. An interesting extension of the model would be to allow for heterogeneous beliefs about the nutrient needs of crops. It would also be interesting to consider extending the NMPs to cover the areas importing manure, and to allow this regulatory extension to influence the willingness to accept manure.

We argue that allowing the farmer to alter the nutrient concentration of manure by feed choices, manure storage, and application techniques would provide qualitatively similar results as the crop choice in the current model. Empirically, however, it would be interesting to analyze how the farmer would optimally increase or decrease the nutrient concentration of manure within the feasible range for each production animal. Increasing nutrient concentration reduces hauling costs of a nutrient unit and makes it more competitive against mineral fertilizers.

References

1. Aillery, M., Gollehon, N., Johansson, R., Kaplan, J., Key, N., Ribaudo, M.: Managing manure to improve air and water quality. Technical Report 9. USDA Economic Research Service (2005a)
2. Aillery, M., Gollehon, N., Breneman, V.E.: Technical documentation of the regional manure management model for the Chesapeake Bay watershed. USDA-ERS Technical Bulletin No. 1913. (2005b)
3. Alexander, R.B., Smith, R.A., Schwarz, G.E., Boyer, E.W., Nolan, J.V., Brakebill, J.W.: Differences in phosphorus and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin. Environ. Sci. Technol. **42**(3), 822–830 (2007)
4. Baerenklau, K., Nergis, N., Schwabe, K.: Effects of nutrient restrictions on confined animal facilities: insights from a structural-dynamic model. Can. J. Agric. Econ. **56**, 219–241 (2008)
5. Baumol, W.J., Oates, W.E.: The Theory of Environmental Policy, 2nd edn. Cambridge (1988)
6. Carpenter, S.R., Caraco, N.F., Correll, D.L., Howarth, R.W., Sharpley, A.N., Smith, V.H.: Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. **8**(3), 559–568 (1998)
7. Carpenter, S.R.: Phosphorus control is critical to mitigating eutrophication. Proc. Natl. Acad. Sci. **105**(32), 11039–11040 (2008)
8. Conley, D., Paerl, H., Howarth, R., Boesch, D., Seitzinger, S., Havens, K., Lancelot, C., Likens, G.: Controlling eutrophication: nitrogen and phosphorus. Science **323**, 1014–1015 (2009)
9. Eghball, B.: Soil properties as influenced by phosphorus-and nitrogen-based manure and compost applications. Agronomy J. **94**(1), 128–135 (2002)
10. Ekholm, P., Lehtoranta, J.: Does control of soil erosion inhibit aquatic eutrophication. J. Environ. Manag. **93**, 140–146 (2012)
11. Feinerman, E., Bosch, D.J., Pease, J.W.: Manure applications and nutrient standards. Am. J. Agric. Econ. **86**(1), 14–25 (2004)
12. Fernandez, L.: Trade's dynamic solutions to transboundary pollution. J. Environ. Econ. Manag. **43**, 386–411 (2002)
13. Fleming, R., Babcock, B., Wang, E.: Resource or waste? the economics of swine manure storage and management. Rev. Agric. Econ. **20**(1), 96–113 (1998)
14. Hesketh, N., Brookes, P.: Development of an indicator for risk of phosphorus leaching. J. Environ. Qual. **29**(1), 105–110 (2000)

15. Hochman, E., Zilberman, D.: Two-goal environmental policy: an integration of micro and macro ad hoc decision rules. J. Environ. Econ. Manag. **6**, 152–174 (1979)
16. Huang, W., Magleby, R., Christensen, L.: Economic impacts of epa's manure application regulations on dairy farms with lagoon liquid systems in the Southwest Region. J. Agric. Appl. Econ. **37**(1), 209–227 (2005)
17. Innes, R.: The economics of livestock waste and its regulation. Am. J. Agric. Econ. **82**(1), 97–117 (2000)
18. Jarvie, H.P., Johnson, L.T., Sharpley, A.N., Smith, D.R., Baker, D.B., Bruulsema, T.W., Confesor, R.: Increased soluble phosphorus loads to Lake Erie: unintended consequences of conservation practices. J. Environ. Qual. **46**(1), 123–132 (2017)
19. Kaplan, J., Johansson, R., Peters, M.: The manure hits the land: economics and environmental implications when land application of nutrients is constrained. Am. J. Agric. Econ. **86**(3), 688–700 (2004)
20. Lake Erie LaMP: Lake Erie binational nutrient management strategy: protecting Lake Erie bymanaging phosphorus. Technical report, Prepared by the Lake Erie LaMP Work Group Nutrient Management Task Group (2011)
21. List, J., Mason, C.: Optimal institutional arrangements for transboundary pollutants in a second-best world: evidence from a differential game with asymmetric players. J. Environ. Econ. Manag. **42**, 277–296 (2001)
22. Maguire, R.O., Kleinman, P.J.A., Dell, C., Beegle, D.B., Brandt, R.C., McGrath, J.M., Ketterings, Q.M.: Manure management in reduced tillage and grassland systems: a review. J. Environ. Qual. **40**, 292–301 (2011)
23. Maler, K.-G., de Zeeuw, A.: The acid rain differential game. Environ. Res. Econ. **12**, 167–184 (1998)
24. Martin, E., Stahn, H.: Potential conflict and inefficiencies arising in agri-environmental management. J. Agric. Econ. **64**(2), 423–445 (2013)
25. Maryland: Title 15 department of agriculture subtitle 20 soil and water conservation chapter08 content and criteria for a nutrient management plan developed for an agricultural operation, Technical report. Authority: Agriculture Article, **8(801)**, 8–806, Annotated Code of Maryland. (2014)
26. McDowell, R., Sharpley, A., Brookes, P., Poulton, P.: Relationship between soil test phosphorus and phosphorus release to solution. Soil Sci. **166**(2), 137–149 (2001)
27. Moffitt, L., Zilberman, D., Just, R.: A "putty-clay" approach to aggregation of production/pollution possibilities: an application in dairy waste control. Am. J. Agric. Econ. **60**, 452–459 (1978)
28. Moslener, U., Requate, T.: Optimal abatement in dynamic multi-pollutant problems when pollutants can be complements or substitutes. J. Econ. Dyn. Control **31**, 2293–2316 (2007)
29. National Research Council: Nutrient control actions for improving water quality in the Mississippi river basin and northern gulf of Mexico. National Academies Press (2009)
30. Oates, W.E.: A Reconsideration of Environmental Federalism, vol. 439. Resources for the Future, Washington, DC. (2001)
31. Pennsylvania: Land application of manure, a supplement to manure management for environmental protection , manure management plan guidance 361-0300-002, Technical report. Pennsylvania Department of Environmental Protection (2011)
32. Pote, D., Daniel, T., Sharpley, A., Moore Jr., P., Edwards, D., Nichols, D.: Relating extractable soil phosphorus to phosphorus losses in runoff. Soil Sci. Soc. Am. J. **60**, 855–859 (1996)
33. Quigg, A., Sylvan, J.B., Gustafson, A.B., Fisher, T.R., Oliver, R.L., Tozzi, S., Ammerman, J.W.: Going west: Nutrient limitation of primary production in the Northern Gulf of Mexico and the importance of the Atchafalaya River. Aquat. Geochem. **17**, 519–544 (2011)
34. Rabalais, N.N., Turner, R.E., Scavia, D.: Beyond science into policy: Gulf of Mexico hypoxia and the Mississippi river. Bio-Science **52**(2), 129–142 (2002)
35. Ribaudo, M., Agapoff, J.: Importance of cost offsets for dairy farms meeting a nutrient application standard. Agric. Resour. Econ. Rev. **34**(2), 173–184 (2005)

36. Ribaudo, M.O., Horan, R.D.: The role of education in nonpoint source pollution control policy. Rev. Agric. Econ. **21**(2), 331–343 (1999)
37. Rougoor, C.W., Van Zeijts, H., Hofreither, M.F., Backman, S.: Experiences with fertilizer taxes in Europe. J. Environ. Plann. Manag. **44**(6), 877–887 (2001)
38. San Jacinto: San Jacinto watershed integrated regional dairy management plan, Prepared for the san jacinto basin resource conservation district, TetraTech. (2009). http://www.waterboards.ca.gov/rwqcb8/waterissues/programs/dairies/docs//1ReportI
39. Scavia, D., Donnelly, K.A.: Reassessing hypoxia forecasts for the Gulf of Mexico. Environ. Sci. Technol. **41**, 8111–8117 (2007)
40. Schindler, D.: The dilemma of controlling cultural eutrophication of lakes. Proc. R. Soc. B **279**, 4322–4333 (2012)
41. Schnitkey, G., Miranda, M.: The impact of pollution controls on livestock-crop producers. J. Agric. Resour. Econ. **18**, 25–36 (1993)
42. Sharpley, A., Daniel, T., Sims, T., Lemunyon, J., Stevens, R.: Agricultural Phosphorus and Eutrophication. Agricultural Research Service, USA (2003)
43. Sharpley, A.N.: Agriculture and Phosphorus Management: The Chesapeake Bay. CRC Press (1999)
44. Sigman, H.: Transboundary spillovers and decentralization of environmental policies. J. Environ. Econ. Manag. **50**, 82–101 (2005)
45. Sims, J., Bergstrom, L., Bowman, B., Oenema, O.: Nutrient management for intensive animal agriculture: policies and practices for sustainability. Soil Use Manag. **21**(1), 141–151 (2005)
46. Smith, K., Jackson, D., Pepper, T.: Nutrient losses by surface run-off following the application of organic manures to arable land. 1. nitrogen. Environ. Pollut. **112**(1), 41–51 (2001)
47. Smith, K., Jackson, D., Withers, P.: Nutrient losses by surface run-off following the application of organic manures to arable land. 2. phosphorus. Environ. Pollut. **112**(1), 53–60 (2001)
48. Smith, V.H.: Low nitrogen to phosphorus ratios favor dominance by blue-green algae in lake phytoplankton. Science(Washington) **221**(4611), 669-671 (1983)
49. Smith, V.H., Joye, S.B., Howarth, R.W.: Eutrophication of freshwater and marine ecosystems. Limnology Oceanogr. **51**(1), 351–355 (2006)
50. Task Force: Action plan for reducing, mitigating, and controlling hypoxia in the Northern Gulf of Mexico. Technical report (2001)
51. Task Force: Gulf Hypoxia Action Plan for reducing, mitigating, and controlling hypoxia in the Northern Gulf ofMexico and improving water quality in the Mississippi river basin. Technical report (2008)
52. Task Force: Reassessment 2013, assessing progress made since 2008. Technical report. (2013)
53. Tietenberg, T.H.: Transferable discharge permits and the control of stationary source air pollution: a survey and synthesis. Land Econ. **56**(4), 391–416 (1980)
54. Tinbergen, J.: On the Theory of Economic Policy. North Holland, Amsterdam (1952)
55. USDA-NRCS.: Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Chesapeake Bay Region (2011)
56. USDA: 2012 census volume 1, chapter 1: U.s. national level data, Technical report, U.S. Department of Agriculture. (2012)
57. Walls, M., Palmer, K.: Upstream pollution, downstream waste disposal, and the design of comprehensive environmental policies. J. Environ. Econ. Manag. **41**, 94–108 (2001)

An Overview of Synchrony in Coupled Cell Networks

Manuela A. D. Aguiar and Ana P. S. Dias

Abstract One of the key aspects in the theory of coupled cell networks concerns the existence of synchrony subspaces. That is, subspaces defined in terms of equalities between cell coordinates which are flow-invariant for all coupled cell systems that respect a given coupled cell network structure. We review some recent concepts and results concerning synchrony subspaces on coupled cell networks. The existence of such subspaces naturally restricts the dynamics that can occur at the coupled cell systems, as in general it is the case for any dynamical system admitting flow-invariant spaces. We focus at some of the aspects that make important and special the existence of synchrony subspaces for coupled cell systems. Namely, their existence depend on the network structure and not on the specific form of the differential equations that are chosen to govern the dynamics; the solutions of the restricted coupled cell systems represent dynamics where groups of cells are dynamically behaving exactly in the same way; the restricted coupled cell systems are again coupled cell systems that are consistent with a network structure with a fewer cells. We review some results on how synchrony changes, or it is combined, in evolving networks. More precisely, in networks where their topology changes with time, either to a rewiring of a link, appearance or removal of a link or a node, or by merging smaller networks into larger ones. Finally, we consider the complement network of a network remarking that both networks have the same set of synchrony subspaces.

Keywords Coupled cell system · Coupled cell network · Synchrony

M. A. D. Aguiar (✉)
Faculdade de Economia, Centro de Matemática, Universidade do Porto,
Rua Dr Roberto Frias, 4200-464 Porto, Portugal
e-mail: maguiar@fep.up.pt

A. P. S. Dias
Faculdade de Ciências, Dep. Matemática, Centro de Matemática,
Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal
e-mail: apdias@fc.up.pt

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_2

1 Introduction

Many real life phenomena can be dynamically modeled through differential equations that can be interpreted as *coupled cell systems* – that is – equations consistent with a *network* graph structure where nodes (the *cells*) symbolize dynamics of smaller dynamical systems and edges represent interactions (the *couplings*) between those nodes. The collective dynamics of the time evolution at nodes then gives the dynamics on the network. In the analysis of the collective dynamics it is often crucial and of interest to observe the dynamical behavior of the individual nodes, comparing and finding features such as synchrony or specified phase-relations in periodic solutions. We follow here the theory of coupled cell networks formalized by Stewart et al. [28, 31, 44] and Field [23]. A key advantage of these formalisms is that it allows theoretical deduction of collective dynamics based only on the network structure, without referring to specific dynamics at every cell.

Different factors can contribute for the decision of modelling through network equations. One such factor can be derived from the intrinsic form of the phenomena that is is being modelled in the mathematical language. As an example, network of symmetrically coupled cells can be used to model central pattern generators for quadruped locomotion, see Golubitsky et al. [17, 29]. We are interested in networks associated with directed graphs meaning that the interactions are directional. For example, in a social network representing trade among nations, the interactions are directional and the graph representing such interactions must be directed. Moreover, many interactions are valued, indicating for example the strength of interaction between the social nodes or there can be more than one type of interaction (multirelational networks). See for example Wasserman and Faust [45]. The theory of coupled cell networks that we are following considers networks represented by directed graphs that can have more than one edge type, multi-edges and self-loops. Graphically, each edge type is represented by a different symbol.

Coupled Cell Networks and Coupled Cell Systems

A network is said to be *regular* if all cells are *identical* (have the same internal dynamics), all edges are of the same type and all cells receive the same number of input edges – the *valency* of the network. More generally, a network such that each subnetwork formed by the network cells and the network edges of a given type is regular, is said to be *homogeneous*. To each such subnetwork is associated an *adjacency matrix*, with rows and columns indexed by the network cells, with nonnegative integers entries, where the entry ij is m if there are m edges of that given type from cell j to cell i. Thus an homogeneous network with k edges types can be described through k (adjacency) matrices.

Example 1 The network N of Fig. 1 is an example of a five-cell regular network of valency two with adjacency matrix

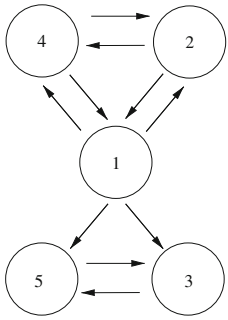


Fig. 1 A five-cell network N which is regular of valency two: all the cells and all the edges are represented by the same cell and edge symbol, respectively, since the cells are identical and the edges are of the same type; all the cells receive two input edges

$$\begin{pmatrix} 0\,1\,0\,1\,0 \\ 1\,0\,0\,1\,0 \\ 1\,0\,0\,0\,1 \\ 1\,1\,0\,0\,0 \\ 1\,0\,1\,0\,0 \end{pmatrix}.$$

The associated coupled cell systems satisfy the following general form:

$$\begin{aligned} \dot{x}_1 &= f(x_1, \overline{x_2, x_4}) \\ \dot{x}_2 &= f(x_2, \overline{x_1, x_4}) \\ \dot{x}_3 &= f(x_3, \overline{x_1, x_5}) \\ \dot{x}_4 &= f(x_4, \overline{x_1, x_2}) \\ \dot{x}_5 &= f(x_5, \overline{x_1, x_3}) \end{aligned}$$

where $f : (\mathbf{R}^k)^3 \to \mathbf{R}^k$, for $k \in \mathbf{N}$, is a smooth function. The overbar indicates that f is invariant under the permutation of the variables and translates the fact that all the interactions (edges) between cells are of the same type. The same function f is used to describe the time evolution of each cell state for two reasons: the same symbol is used to represent all the cells which indicates that the cells are identical; each cell receives two interactions and so the equations for each cell are identical up to the input variables.

When studying the dynamics of coupled cell systems, obviously it has to be taken into account the underlined network structure, which in particular, can force dynamics that would be highly nongeneric in the context of general dynamical systems. One such example is the occurrence of flow-invariant spaces.

Synchrony Subspaces

One widely observed and most studied collective dynamics in coupled dynamical systems is the *synchronization*, where phase trajectories of two or more coupled units

coincide over time. For importance of synchronization and its ubiquitous presence in nature, we refer to [15, 41] and references therein. In [40], Pikovsky et al. propose to study various synchronization phenomena using a common framework based on modern nonlinear dynamics, where a variety of approaches using coupled periodic and coupled chaotic systems is discussed. Restrepo et al. in [42] point out the crucial effect of network structure on the emergence of collective synchronization in heterogeneous systems, in terms of eigenvalues of network adjacency matrices.

Conditions for the occurrence of robust patterns of partial synchronization in terms of network structure, have been established in Stewart et al. [44] and Golubitsky et al. [31]. In the theory of coupled cell networks the synchronization of two or more cells corresponds to the flow-invariance of the subspace of the total phase space given by the identification of the phase space of those cells. These are called *synchrony subspaces* and have the amazing property that their existence, implying flow-invariance for the associated coupled cell systems, depends only on the network structure. In fact, by [31, 44] synchrony subspaces are in one-to-one correspondence with the equivalence relations on the network set of cells that satisfy certain properties in which case they are called balanced. Equivalently, synchrony subspaces are in one-to-one correspondence with the polydiagonals (subspaces of $\mathbf{R}^n$, if the original network has n cells, defined by equalities of coordinates) that are left invariant under the network adjacency matrix, or the adjacency matrices if the network has more than one edge type.

Example 2 Consider the five-cell regular network N of Fig. 1 with set of cells $C = \{1, \ldots, 5\}$ and total phase space $(\mathbf{R}^k)^5$. The polydiagonal subspace $\Delta = \{\mathbf{x} \in (\mathbf{R}^k)^5 : \ x_1 = x_2, \ x_3 = x_5\}$ is a synchrony subspace for the coupled cell systems associated to N. This is easily verified using the general form of the equations of the admissible vector fields for N, presented in Example 1. With the identification of x_1 with x_2 and of x_3 with x_5, the equations for $\dot{x}_1$ and $\dot{x}_2$ coincide and the equations for $\dot{x}_3$ and $\dot{x}_5$ also coincide. Thus a trajectory with initial condition in Δ will remain in Δ for all time. Equivalently, from the results of [31, 44], Δ is a synchrony subspace since the polydiagonal subspace $\{\mathbf{x} \in \mathbf{R}^5 : \ x_1 = x_2, \ x_3 = x_5\}$ is left invariant under the adjacency matrix of N, presented in Example 1.

Symmetry and Synchrony

Symmetric networks are a special class of networks. The *symmetry group* of a network is the group of the isomorphisms of the network graph. Equivalently, the symmetry group of the network corresponds to the group of the $n \times n$ permutation matrices (if the network has n cells) that commute with the adjacency matrix or the adjacency matrices of the network. Coupled cell systems associated with symmetric networks inherit the network symmetry – that is – they are equivariant under the network symmetry group, considering the natural action by permutation of the network cell coordinates. In this case, there are two main aspects that determine the form of

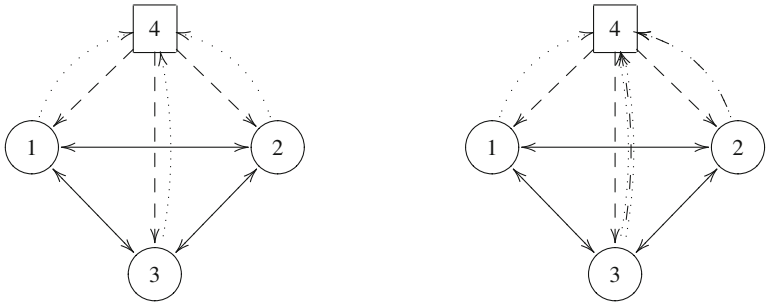


Fig. 2 (left) Network with exact $\mathbf{S}_3$-symmetry. (right) Network with $\mathbf{S}_3$-interior symmetry on the set of cells $\{1, 2, 3\}$

the coupled cell systems - the network and the symmetry. That is, the coupled cell systems are equivariant under the network symmetry group and are also constrained by the network structure. For any isotropy subgroup for the action of the network group of symmetries, the corresponding fixed-point subspace is flow-invariant and it is a polydiagonal, since the action is by permutation of the network coordinates, thus it is a synchrony subspace. But, there can be more additional synchrony subspaces whose existence is not predicted by the symmetry. This comes from the fact that the coupled cell systems are not only equivariant but they also have form consistent with the network. More precisely, the linear space of smooth vector fields with structure consistent with the symmetric network may form a proper subspace of the linear space of the smooth equivariant vector fields. See Antoneli and Stewart [12–14]. It is then possible that dynamics that are non-generic from the symmetric point of view, are generic for a given symmetric network structure. That is, dynamics can occur in a robust way for coupled cell systems that have form consistent with a specific network structure, but that would not be expected if we were working in the context of generic smooth equivariant vector fields. See for example Golubitsky et al. [25]. See also Dias and Lamb [19], Paiva [39, Chap. 7], Dias and Paiva [21] and Golubitsky and Lauterbach [24].

An important class of non-symmetric networks that lies between the class of general networks and the class of symmetric networks, where group theoretic methods still apply, are the networks with *interior symmetries*. In this case, there is a group of permutations of a subset S of the cells (and edges directed to S) that partially preserves the network structure (including cell-types and edges-types) and its action is again by permutation of the network cell coordinates. In other words, the cells in S together with all the edges directed to them form a subnetwork which possesses a non-trivial group of symmetry $\Sigma_S \subseteq \mathbf{S}_n$. For example, in Fig. 2, the network at the left has exact $\mathbf{S}_3$-symmetry, whereas the network on the right has $\mathbf{S}_3$-interior symmetry on the set of cells $S = \{1, 2, 3\}$. This notion was introduced and investigated by Golubitsky, Pivato and Stewart [26].

In coupled cell systems, the local bifurcations from a synchronous equilibrium can be classified into *synchrony-breaking* bifurcations or *synchrony-preserving*

bifurcations. The synchrony-breaking bifurcations occur when a synchronous state looses stability and bifurcates to a state with less synchrony. This is in parallel with the concept of symmetry-breaking bifurcations in symmetric coupled cell systems, see Golubitsky and Stewart [27]. In [26] it is obtained analogues of the Equivariant Branching Lemma [30, Theorem XIII 3.3] and the Equivariant Hopf Theorem [30, Theorem XVI 4.1] for coupled cell systems with interior symmetries.The analogue of the Equivariant Branching Lemma is a natural generalization of the symmetric case. However, in the Equivariant Hopf Theorem, it is proved the existence of states whose linearizations on certain subsets of cells, near bifurcation, are superpositions of synchronous states with states having 'spatial symmetries'. (In the full symmetric case, the Equivariant Hopf Theorem guarantees the existence of states with certain spatio-temporal symmetries.) More recently, in Antoneli, Dias and Paiva [10, Theorem 4.8], the Equivariant Hopf Theorem for networks with interior symmetries of [26] is extended obtaining the full analogue of the Equivariant Hopf Theorem for networks with symmetries. More precisely, it is guaranteed the existence of states whose *linearizations* on certain subsets of cells, near bifurcation, are superpositions of synchronous states with states having *spatio-temporal symmetries*, that is, corresponding to "interiorly" $\mathbf{C}$-axial subgroups of $\Sigma_S \times \mathbf{S}^1$. See also Antoneli, Dias and Paiva [11].

Applying the Equivariant Hopf Theorem to a smooth one-parameter family of coupled cell systems with structure consistent with the network at the left of Fig. 2 which has exact $\mathbf{S}_3$-symmetry, assuming a codimension-one interior symmetry-breaking Hopf bifurcation occurs at an equilibrium with $\mathbf{S}_3$-symmetry, then generically we obtain three branches of small amplitude periodic solutions. One branch corresponds to periodic solutions with exact spatial $\mathbf{Z}_2$-symmetry where two cells undergo oscillations that are identical and in phase, and the third (from the set $\{1, 2, 3\}$) behaving differently. There are two more branches of periodic solutions with spatio-temporal symmetries $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2$: on one branch the oscillations have the same waveform for each cell in the set $\{1, 2, 3\}$, but are phase-shifted by one third of the period; at the other branch, two cells have identical waveforms but are one half of the period out of phase, and the third cell (from the set $\{1, 2, 3\}$) has the double frequency. The three groups $\mathbf{Z}_2$, $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2$ correspond to the three (conjugacy classes of) isotropy subgroups of the standard action of $\mathbf{S}_3 \times \mathbf{S}^1$ on $\mathbf{C} \oplus \mathbf{C}$. For details see for example Golubitsky, Stewart and Schaeffer [30, Chaps. XVI, XVIII]. Applying the Equivariant Hopf Theorem with Interior Symmetries of [10], now taking a smooth one-parameter family of coupled cell systems with structure consistent with the network at the right of Fig. 2 which has interior $\mathbf{S}_3$-symmetry on the set of cells $S = \{1, 2, 3\}$, assuming a codimension-one interior symmetry-breaking Hopf bifurcation occurs at an equilibrium with $\mathbf{S}_3$-symmetry, then generically we obtain as well three branches of small amplitude periodic solutions, corresponding to the three groups $\mathbf{Z}_2$, $\tilde{\mathbf{Z}}_3$ and $\tilde{\mathbf{Z}}_2$, but now the linearizations of the periodic states on the subsets of cells, near bifurcation, are superpositions of synchronous states with states having those spatio-temporal symmetries. See the numerical simulations in [10, Sect. 4.4] illustrating the periodic solutions guaranteed by the Equivariant Hopf Theorem with exact and interior $\mathbf{S}_3$-symmetry in coupled cell systems of four cells with structure consistent with the

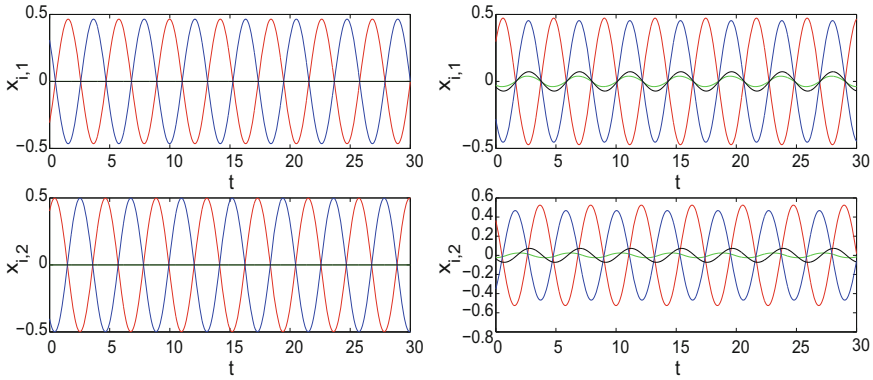


Fig. 3 Solutions with $\tilde{\mathbf{Z}}_2$ (interior) symmetry. (left) Network with exact $\mathbf{S}_3$-symmetry. (right) Network with $\mathbf{S}_3$-interior symmetry. Figure taken from [10]

networks of Fig. 2, respectively, choosing the internal phase space of all four cells to be $\mathbf{C} \cong \mathbf{R}^2$. We reproduce here in Fig. 3 results of numerical simulations obtaining periodic solutions with (interior) $\tilde{\mathbf{Z}}_2$-symmetry: it is superimposed the time series of all four cells, which are identified by colours: cell 1 is blue, cell 2 is red, cell 3 is green, and cell 4 is black. The upper panels show the first components and the lower panels show the second components. The left panels refer to network with exact $\mathbf{S}_3$-symmetry and the panels on the right refer to network with $\mathbf{S}_3$-interior symmetry.

Quotients and Inflations

Synchrony subspaces have a major impact at the dynamics of the coupled cell systems associated with a given network. An important aspect of the existence of synchrony subspaces is that the restriction of the coupled cell systems to a synchrony subspace are again coupled cell systems in a lower-dimensional phase space, now associated with a network with fewer cells – the *quotient network* of the given network by the synchrony subspace. The fact that the restricted systems are consistent with a network structure implies constrains at the dynamics that can occur for those systems and thus for the initial coupled cell systems. Although, the restrictions to the synchrony subspaces do not give all the dynamics for the original network, they give full information concerning the dynamics of the original coupled cell systems at those synchrony subspaces. See for example Aguiar et al. [4, 5]. Moreover, it can happen that the quotient network has been already explored from the dynamical point of view in several contexts. If that is the case, the known dynamics of the quotient network can be lifted to the original network dynamics. Examples of specific structures than can be explored are: existence of global (quotient) network symmetries implying that the associated coupled cell systems are symmetric under a permutation symmetry group

– these impose strong constrains at the dynamics that can occur, see for example Golubitsky and Stewart [27, 28] and references therein; known classifications of classes of networks with certain structures, see for example Leite and Golubitsky [35].

It is known that flow-invariant spaces favour the existence of non-generic dynamical behaviour like heteroclinic cycles and networks, which lead to complicated dynamics. It follows that, knowing the set of all synchrony subspaces of a coupled cell network, can help to detect the possibility of the associated coupled cell systems to support heteroclinic behaviour. Besides this, in Aguiar et al. [1], it is also explored the process reverse to the quotient of a network: coupled cell networks supporting heteroclinic networks are constructed by lifting coupled cell dynamics supporting heteroclinic behaviour and associated with smaller networks. That is, networks with a few number of cells (and supporting heteroclinic connections) are *inflated* and combined in a way that they are quotient networks of bigger networks supporting heteroclinic networks.

Example 3 In Fig. 4 we show a six-cell network M for which the five-cell network N of Fig. 1 is a quotient network by the synchrony subspace defined by the cell coordinates equality $x_1 = x_6$. More precisely, the general form of the coupled cell systems associated with M is:

$$\begin{aligned}
\dot{x}_1 &= f(x_1, \overline{x_2, x_4}) \\
\dot{x}_2 &= f(x_2, \overline{x_6, x_4}) \\
\dot{x}_3 &= f(x_3, \overline{x_6, x_5}) \\
\dot{x}_4 &= f(x_4, \overline{x_1, x_2}) \\
\dot{x}_5 &= f(x_5, \overline{x_1, x_3}) \\
\dot{x}_6 &= f(x_6, \overline{x_2, x_4})
\end{aligned}$$

where as before, $f : (\mathbf{R}^k)^3 \to \mathbf{R}^k$ is any smooth function and the overbar indicates that f is invariant under the permutation of the variables. Restricting these equations to the synchrony subspace $\{\mathbf{x} : x_1 = x_6\}$, we obtain the general form of the coupled cell systems associated with the network N of Fig. 1 (see Example 1). Equivalently, the network M in Fig. 4 is an inflation of the five-cell network N of Fig. 1 at cell 1, where cell 1 is inflated to cells 1 and 6.

2 Enumeration of Inflations

In general, a given network can be the quotient network of many different networks. In Aguiar et al. [4, 5] it is considered the inverse problem: given a network N provide a systematic way of enumerating the networks that admit N as a quotient network. Those networks are called *inflations* (Aguiar et al. [1]) or *lifts* (Dias and Moreira [20]) of N.

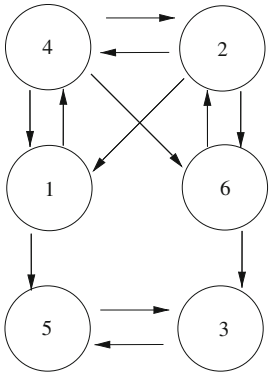


Fig. 4 A network M which has the network N of Fig. 1 as a quotient by the synchrony subspace defined by the cell coordinates equality $x_1 = x_6$. We also say that the six-cell network M is a simple inflation of the network N of Fig. 1 at cell 1, where cell 1 is inflated to cells 1 and 6

2.1 Inflating (Lifting) a Network

An inflation (lift) M of N can be interpreted as enlarging the network N in the number of cells, preserving the valency, where each cell of N corresponds to the identification of a certain set of cells in the inflation M. In order that an n-cell network M is a lift of N, given any two cells that were identified, they must receive the same number of directed edges from cells of each class of identified cells.

An inflation is said to be a *simple inflation* if there is just one cell that is inflated.

Example 4 If we take the five-cell regular network in Fig. 1, then one of its six-cell (simple) inflations is the network in Fig. 4 where cell 1 of N is inflated to cells 1 and 6 of M. Observe that in N, cell 1 receives two directed edges, one from cell 4 and one from cell 2, and sends two directed edges to cells 5 and 3. The network in Fig. 4 is an inflation of N since: cells 1 and 6, both receive one directed edge from each of the cells 4 and 2; there is a directed edge from one of the cells in the class $\{1, 6\}$ to both cells 5 and 3 – a directed edge from cell 1 to cell 5 and a directed edge from cell 6 to cell 3.

Using the theory of coupled cell networks [31, 44], one way to enumerate all the possible inflation networks M, for a fixed N and a fixed polydiagonal, is through the construction of the possible $n \times n$ (adjacency) matrices leaving invariant the fixed polydiagonal and whose restrictions to the polydiagonal are similar to the adjacency matrix of the network N. The methods of enumeration presented by Aguiar et al. [4, 5] explore precisely this approach and are developed for regular networks. (See also Dias and Moreira [20].) In fact, these methods trivially extend to homogeneous networks. Recall that an homogeneous network can be seen as a directed graph with more than one edge type, where the subnetworks on the same network set of cells, considered for each edge-type, are regular. Finding the set of synchrony subspaces of an homogeneous network is equivalent to find the common synchrony subspaces of all these subnetworks. Moreover, if the network is not homogeneous, now the subnetworks to be considered are in some way homogeneous and then the question

is again reduced to consider homogeneous networks, and then, regular networks, see Aguiar and Dias [2].

2.2 *Inflating (Lifting) a Bifurcation*

Consider a coupled cell system with structure consistent with a regular network M, depending on a real bifurcation parameter and assume that a codimension-one steady-state or Hopf bifurcation occurs at a full synchronous equilibrium X_0 which, after an affine change of coordinates, we can assume is the null steady-state solution X_0. Note that, the full diagonal space is always a synchrony subspace of a regular network. In Aguiar et al. [5] it is addressed the problem of how a steady-state or Hopf bifurcation occurring at a quotient network N of M lifts to M. Every bifurcating solution for the quotient lifts to a bifurcation solution for the inflation network where cells that were identified in the quotient are synchronized. But it can occur that new bifurcating solutions appear for the inflation network M where cells that were identified in the quotient are not synchronized. In Aguiar et al. [5], examples are given of five-cell networks with the three-cell bidirectional ring as quotient, where bifurcations within the ring dynamics lead to solutions that break synchrony in the five-cell network. One of those is the network of Fig. 1.

Results in Leite and Golubitsky [35] and Golubitsky and Lauterbach [24] relate the eigenvalues of the Jacobian J_M of a coupled cell system consistent with a network M at X_0 with the eigenvalues of the adjacency matrix of M. In order for bifurcations within the quotient network N to lead to nonsynchronous solutions in the larger network M, the center subspace of J_M must be larger than the center subspace of J_N. Results are presented in [5] that relate the eigenvalues of the adjacency matrix of the network M with those of the adjacency matrix of the quotient N which provide an easy way to identify networks M for which the dimension of the center subspaces of J_M and J_N are the same. Each one-parameter steady-state (or Hopf) bifurcation supported by the coupled cell systems for M (or for N) is associated with a degeneracy condition corresponding to a zero (or imaginary) eigenvalue of J_M (or J_N) that depends at the eigenvalues of the adjacency matrix of M (or N). The eigenvalues of the adjacency matrix of any inflation M of N are the eigenvalues of N plus other eigenvalues, following the terminology [20], the *extra* eigenvalues. A degeneracy condition implying a steady-state or Hopf bifurcation of M (or N) is associated at least with an eigenvalue of the adjacency matrix of the network M (or N). That is, the critical eigenvalues of J_M (or J_N) are directly associated with the eigenvalues of the adjacency matrix of M (or N). It is then easy to see that if the real parts of the extra eigenvalues of the adjacency matrix of an inflation M of N are distinct from the real parts of the 'critical' eigenvalues of the adjacency matrix of N then the coupled cell systems associated with the inflation network M will have no additional branches of steady-state solutions (or periodic solutions in the Hopf case), for the fixed imposed bifurcation degeneracy condition. As an example, in [5] it is proved that, up to isomorphism, there are two four-cell and twelve five-cell networks admitting the

three-cell bidirectional ring quotient network, and from these it is shown that only two such networks can exhibit branches of steady-state solutions not predicted by bifurcation in the three-cell bidirectional ring. In fact, generically, the coupled cell systems associated with these two networks have additional branches, and one of these two networks is precisely the network in Fig. 1. More recently some progress has been achieved at this problem. See Dias and Moreira [20] and Moreira [37].

3 The Lattice of Synchrony Subspaces of a Network

As mentioned above, following Stewart et al. [44] and Golubitsky et al. [31], a synchrony subspace of a network is a subspace given by the identification of the phase space of groups of cells (polydiagonal) that is left invariant under any coupled cell system that has form consistent with the network.

By Stewart [43] (see also Aldis [9]) the set of synchrony subspaces associated with a network, taking the relation of inclusion $\subseteq$, is a complete lattice. Recall that a *lattice* is a partially ordered set such that every pair of elements has a unique least upper bound or *join*, and a unique greatest lower bound or *meet*. Moreover, a *complete lattice* X is a lattice where every subset $Y \subseteq X$ has a unique least upper bound or join, and a unique greatest lower bound or meet. We remark that for a regular network there are always two *trivial synchrony subspaces*, the total asynchronous and the full synchronous polydiagonal subspaces, corresponding, respectively, to the top and bottom elements of the lattice.

In [2], Aguiar and Dias describe how to obtain the lattice of synchrony subspaces of a given network. As shown, this reduces basically to the problem of how to obtain the lattice of synchrony subspaces of regular networks, and more generally, to identical-cell identical-edge coupled networks. For a regular network the lattice of synchrony subspaces is obtained based on the eigenvalue structure of the network adjacency matrix. It is presented an algorithm that generates the lattice of synchrony subspaces for a regular network. See also the work of Kamei [33], on the class of regular networks where the adjacency matrix has only simple eigenvalues, Kamei and Cock [34] for a computer algorithm searching for all possible balanced equivalence relations using symbolic matrix computations, and Moreira [38] where the lattice of synchrony subspaces of a regular network is obtained using a special class of Jordan subspaces of the network adjacency matrix.

Example 5 The lattice of synchrony subspaces of the network in Fig. 1 was obtained by running the algorithm presented in [2]. The nontrivial synchrony subspaces are listed in Table 1. The trivial synchrony subspaces for the network are the total asynchronous polydiagonal space and full synchronous polydiagonal space $\{\mathbf{x} : x_1 = x_2 = x_3 = x_4 = x_5\}$, that we will represent by P and Δ_0, respectively. A representation of the lattice is presented in Fig. 5.

Table 1 Nontrivial synchrony subspaces for the network of Fig. 1. The trivial synchrony subspaces for the network are the total asynchronous polydiagonal space and the full synchronous polydiagonal space $\{\mathbf{x} : x_1 = x_2 = x_3 = x_4 = x_5\}$

	$\Delta_5 = \{\mathbf{x} : x_1 = x_2 = x_4\}$	$\Delta_{11} = \{\mathbf{x} : x_1 = x_2 = x_3,\ x_4 = x_5\}$
$\Delta_1 = \{\mathbf{x} : x_1 = x_2\}$	$\Delta_6 = \{\mathbf{x} : x_2 = x_3,\ x_4 = x_5\}$	$\Delta_{12} = \{\mathbf{x} : x_1 = x_2 = x_4,\ x_3 = x_5\}$
$\Delta_2 = \{\mathbf{x} : x_1 = x_4\}$	$\Delta_7 = \{\mathbf{x} : x_1 = x_2,\ x_3 = x_5\}$	$\Delta_{13} = \{\mathbf{x} : x_1 = x_2 = x_5,\ x_3 = x_4\}$
$\Delta_3 = \{\mathbf{x} : x_2 = x_4\}$	$\Delta_8 = \{\mathbf{x} : x_2 = x_4,\ x_3 = x_5\}$	$\Delta_{14} = \{\mathbf{x} : x_1 = x_3 = x_4,\ x_2 = x_5\}$
$\Delta_4 = \{\mathbf{x} : x_3 = x_5\}$	$\Delta_9 = \{\mathbf{x} : x_2 = x_5,\ x_3 = x_4\}$	$\Delta_{15} = \{\mathbf{x} : x_1 = x_4 = x_5,\ x_2 = x_3\}$
	$\Delta_{10} = \{\mathbf{x} : x_1 = x_4,\ x_3 = x_5\}$	$\Delta_{16} = \{\mathbf{x} : x_2 = x_3 = x_4 = x_5\}$

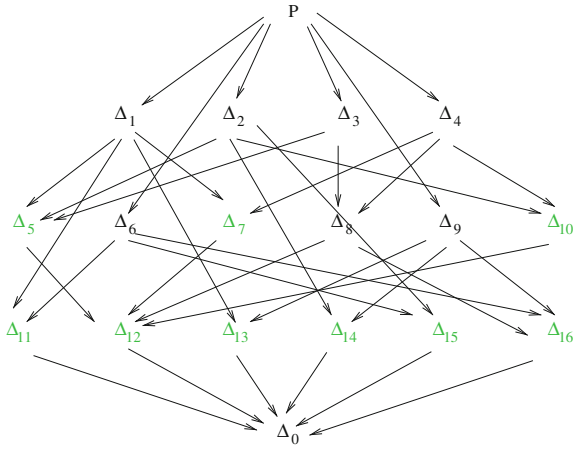


Fig. 5 The lattice of synchrony subspaces for the five-cell regular network N of Fig. 1: the nontrivial synchrony subspaces Δ_i, for $i = 1, \ldots, 16$, are listed in Table 1. The top element is the total phase space P (the total asynchronous polydiagonal space) and the bottom element Δ_0 is the full synchronous polydiagonal space

4 Evolution of Synchrony

Most real world networks are *evolving networks*, that is, their topology evolves with time, either due to a rewiring of a link, the appearance or disappearance of a link or node, or by a merging of small networks into a larger one. The dynamics of network topology reflects frequent changes in the interactions among network components and translates into a rich variety of evolutionary patterns. Evolution of network topology can be described by a sequence of static networks and the topology of the networks can be regarded as a discrete dynamical system. Evolving networks are ubiquitous in nature and science. See Albert et al. [8] and Dorogovtsev et al. [22], and references therein for examples in many diverse fields.

For the different definitions of synchronization, there is a vast literature on how synchronizability varies with the changes of the network structure. As examples, we refer to the works of Atay and Biyikoglu [16], Chen and Duan [18], Lu et al. [36], Hagberg and Schult [32].

In the context of coupled cell systems, since, as mentioned earlier, the connecting topology of a network dictates the lattice of synchrony subspaces, we expect changes

at the corresponding lattice, if the underlying topology changes. In this perspective, the work in Aguiar et al. [6] considers structural changes in the network topology caused by unary network operations, such as deletion and addition of cells or edges, and rewirings of edges, describing which synchrony subspaces are inherited by the new network structure. The works in Aguiar and Ruan [7] and Aguiar and Dias [3] focus on evolving networks where new networks are formed by combining existing ones using binary network operations - the join and the coalescence operations and the direct and tensor product operations, respectively. Results are obtained relating the set of synchrony subspaces of the component networks and the resulting network.

4.1 Inflation

Equivalently to the definition seen before, an inflation (or lift) of a k-cell network N is any network M with $n > k$ cells such that M admits a synchrony subspace where each coupled cell system associated with M restricted to the synchrony subspace is a coupled cell system now consistent with the network N.

Example 6 The network M in Fig. 4 is a six-cell (simple) inflation of the five-cell network in Fig. 1, where cell 1 of N is inflated to cells 1 and 6 of M. From the definition of inflation, it follows that $\tilde{\Delta}_0 = \{\mathbf{x} : \ x_1 = x_6\}$ is a synchrony subspace of M. Moreover, it follows also that there is a one-to-one correspondence between the synchrony subspaces of network N and the synchrony subspaces of network M that are contained in $\tilde{\Delta}_0$. More concretely, for every nontrivial synchrony subspace $\Delta_i, i = 1, \ldots, 16$, for N (recall Table 1), the subspace $\tilde{\Delta}_i$, defined by the coordinate equality conditions that define Δ_i together with the coordinate equality condition $x_1 = x_6$, is a synchrony subspace of M. The nontrivial synchrony subspaces of M are listed in Table 2.

Table 2 Nontrivial synchrony subspaces of the network M of Fig. 4. The network M is an inflation of the five-cell network N of Fig. 1 where cell 1 is inflated to cells 1 and 6. The synchrony subspaces of the network M that are contained in the synchrony subspace $\tilde{\Delta}_0$ are in one-to-one correspondence with the synchrony subspaces of the network N that are listed in Table 1

$\tilde{\Delta}_0 = \{\mathbf{x} : \ x_1 = x_6\}$	$\tilde{\Delta}_5 = \{\mathbf{x} : \ x_1 = x_2 = x_4 = x_6\}$	$\tilde{\Delta}_{11} = \{\mathbf{x} : \ x_1 = x_2 = x_3 = x_6, \ x_4 = x_5\}$
$\tilde{\Delta}_1 = \{\mathbf{x} : \ x_1 = x_2 = x_6\}$	$\tilde{\Delta}_6 = \{\mathbf{x} : \ x_1 = x_6, \ x_2 = x_3, \ x_4 = x_5\}$	$\tilde{\Delta}_{12} = \{\mathbf{x} : \ x_1 = x_2 = x_4 = x_6, \ x_3 = x_5\}$
$\tilde{\Delta}_2 = \{\mathbf{x} : \ x_1 = x_4 = x_6\}$	$\tilde{\Delta}_7 = \{\mathbf{x} : \ x_1 = x_2 = x_6, \ x_3 = x_5\}$	$\tilde{\Delta}_{13} = \{\mathbf{x} : \ x_1 = x_2 = x_5 = x_6, \ x_3 = x_4\}$
$\tilde{\Delta}_3 = \{\mathbf{x} : \ x_1 = x_6, \ x_2 = x_4\}$	$\tilde{\Delta}_8 = \{\mathbf{x} : \ x_1 = x_6, \ x_2 = x_4, \ x_3 = x_5\}$	$\tilde{\Delta}_{14} = \{\mathbf{x} : \ x_1 = x_3 = x_4 = x_6, \ x_2 = x_5\}$
$\tilde{\Delta}_4 = \{\mathbf{x} : \ x_1 = x_6, \ x_3 = x_5\}$	$\tilde{\Delta}_9 = \{\mathbf{x} : \ x_1 = x_6, \ x_2 = x_5, \ x_3 = x_4\}$	$\tilde{\Delta}_{15} = \{\mathbf{x} : \ x_1 = x_4 = x_5 = x_6, \ x_2 = x_3\}$
$\tilde{\Delta}_{17} = \{\mathbf{x} : \ x_1 = x_4\}$	$\tilde{\Delta}_{10} = \{\mathbf{x} : \ x_1 = x_4 = x_6, \ x_3 = x_5\}$	$\tilde{\Delta}_{16} = \{\mathbf{x} : \ x_1 = x_6, \ x_2 = x_3 = x_4 = x_5\}$
$\tilde{\Delta}_{18} = \{\mathbf{x} : \ x_2 = x_6\}$	$\tilde{\Delta}_{19} = \{\mathbf{x} : \ x_1 = x_4, \ x_2 = x_6\}$	$\tilde{\Delta}_{21} = \{\mathbf{x} : \ x_1 = x_4 = x_5, \ x_2 = x_3\}$
	$\tilde{\Delta}_{20} = \{\mathbf{x} : \ x_2 = x_3, \ x_4 = x_5\}$	$\tilde{\Delta}_{22} = \{\mathbf{x} : \ x_2 = x_3 = x_6, \ x_4 = x_5\}$
		$\tilde{\Delta}_{23} = \{\mathbf{x} : \ x_1 = x_4 = x_5, \ x_2 = x_3 = x_6\}$

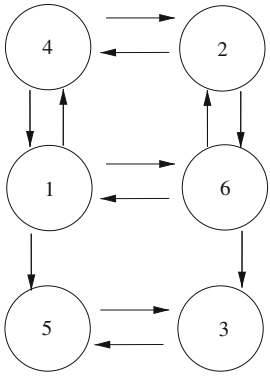


Fig. 6 Network R is a rewiring of network M of Fig. 4, where the directed edges, from cell 2 to cell 1 and from cell 4 to cell 6, are replaced by the directed edges, from cell 6 to cell 1 and from cell 1 to cell 6, respectively

4.2 Rewiring

A *rewiring* of a network occurs when at least one edge of a network is replaced by another edge of the same type and with the same head cell.

Let R be the network obtained by rewiring a edge of a network N. Suppose the rewiring operation replaces an input edge to a cell c from a cell d with one input edge from a cell a. By Lemma 3.10 in Aguiar et al. [6], a polydiagonal Δ is simultaneously a synchrony subspace of M and R if and only if, in the definition of Δ, either there is the coordinate equality condition $d = a$ or there is no coordinate equality condition involving c.

Next we present an example with a rewiring of multiple edges.

Example 7 The network R of Fig. 6 is a rewiring of network M of Fig. 4, where the directed edges, from cell 2 to cell 1 and from cell 4 to cell 6, are replaced by the directed edges, from cell 6 to cell 1 and from cell 1 to cell 6, respectively. It follows from Lemma 3.23 of [6] that the synchrony subspaces of R that are inherited from M are such that in their definition one of the following three conditions holds:

- there is no coordinate equality condition involving x_1 and x_6;
- the only coordinate equality condition involving x_1 and x_6 is $x_1 = x_6$, and for all $i \neq 1, 6$ and $j \in \{2, 4\}$, if there is the coordinate equality condition $x_j = x_i$ then there is also the coordinate equality condition $x_k = x_i$ for all $k \in \{2, 4\} \setminus \{j\}$;
- for all i and for all $j \in \{1, 4\}$ if there is the coordinate equality condition $x_j = x_i$ then there is also the coordinate equality condition $x_k = x_i$ for all $k \in \{1, 4\} \setminus \{j\}$. Moreover, for all i and for all $j \in \{2, 6\}$ if there is the coordinate equality condition $x_j = x_i$ then there is also the coordinate equality condition $x_k = x_i$ for all $k \in \{2, 6\} \setminus \{j\}$.

We have then that the synchrony subspaces for R that are inherited from M are the synchrony subspaces $\tilde{\Delta}_3$, $\tilde{\Delta}_5$, $\tilde{\Delta}_8$, $\tilde{\Delta}_{12}$, $\tilde{\Delta}_{16}$, $\tilde{\Delta}_{19}$, $\tilde{\Delta}_{20}$ and $\tilde{\Delta}_{23}$ from Table 2.

4.3 Product

In [3], Aguiar and Dias consider two product operations on identical-edge networks: the cartesian and the Kronecker (tensor) product.

Definition 1 Let N_1 and N_2 be two identical-edge networks. Assume that N_i has set of cells $C_i = \{1, \ldots, r_i\}$ and set of arrows E_i, for $i = 1, 2$. Consider the cartesian product $C_1 \times C_2$ and denote by ij the element (i, j) in $C_1 \times C_2$.
(i) The *cartesian product* of N_1 and N_2, denoted by $N_1 \boxtimes N_2$, is the network with set of cells $C_1 \times C_2$ and two edge types such that there is an edge from cell ij to cell kl if and only if:

$$i = k \text{ and } (j, l) \in E_2, \text{ or } j = l \text{ and } (i, k) \in E_1 . \tag{1}$$

The edge type of the edges from cells ij to cells il are of distinct type of the edge type of the edges from cells ij to cell kj.
(ii) The *Kronecker product* of N_1 and N_2, denoted by $N_1 \otimes N_2$, is the network with set of cells $C_1 \times C_2$ and such that there is an arrow from cell ij to cell kl if and only if:

$$(i, k) \in E_1 \text{ and } (j, l) \in E_2 . \tag{2}$$

See Fig. 7 for an example of two networks N_1 and N_2, and the product networks $N_1 \boxtimes N_2$, $N_1 \otimes N_2$.

The results in [3] establish an inclusion relation between the lattices of synchrony subspaces for the cartesian and Kronecker products of networks. Specifically, it is proved, in Proposition 4.5, that, for any two identical-edge networks N_1 and N_2, every synchrony subspace of the cartesian product $N_1 \boxtimes N_2$ is a synchrony subspace of the Kronecker product $N_1 \otimes N_2$. For the case of *regular synchrony subspaces*, that is synchrony subspaces of the tensor product $P_1 \otimes P_2$, of the total phase spaces P_1 and P_2 of N_1 and N_2, respectively, of the form $S_1 \otimes S_2$, with S_i a synchrony subspace of P_i, $i = 1, 2$, the results in [3] show equality. That is, the lattice of the regular synchrony subspaces of $N_1 \boxtimes N_2$ is the lattice of the regular synchrony subspaces of $N_1 \otimes N_2$.

Moreover, in [3] it is shown how to obtain the lattice of regular synchrony subspaces of a product network from the lattices of synchrony subspaces of the component networks. Specifically, it is proved that a tensor of subspaces is of synchrony of the product network if and only if the subspaces involved in the tensor are synchrony subspaces for the component networks of the product. It is also shown that, in general, there are (irregular) synchrony subspaces for the product network that are not described by the synchrony subspaces for the component networks, concluding that, in general, it is not possible to obtain the all synchrony lattice for the product network from the corresponding lattices for the component networks.

Example 8 The network R of Fig. 6 is the cartesian product of networks R_1 and R_2 of Fig. 8, that is, $R = R_1 \boxtimes R_2$. (Here we are assuming a slightly different definition

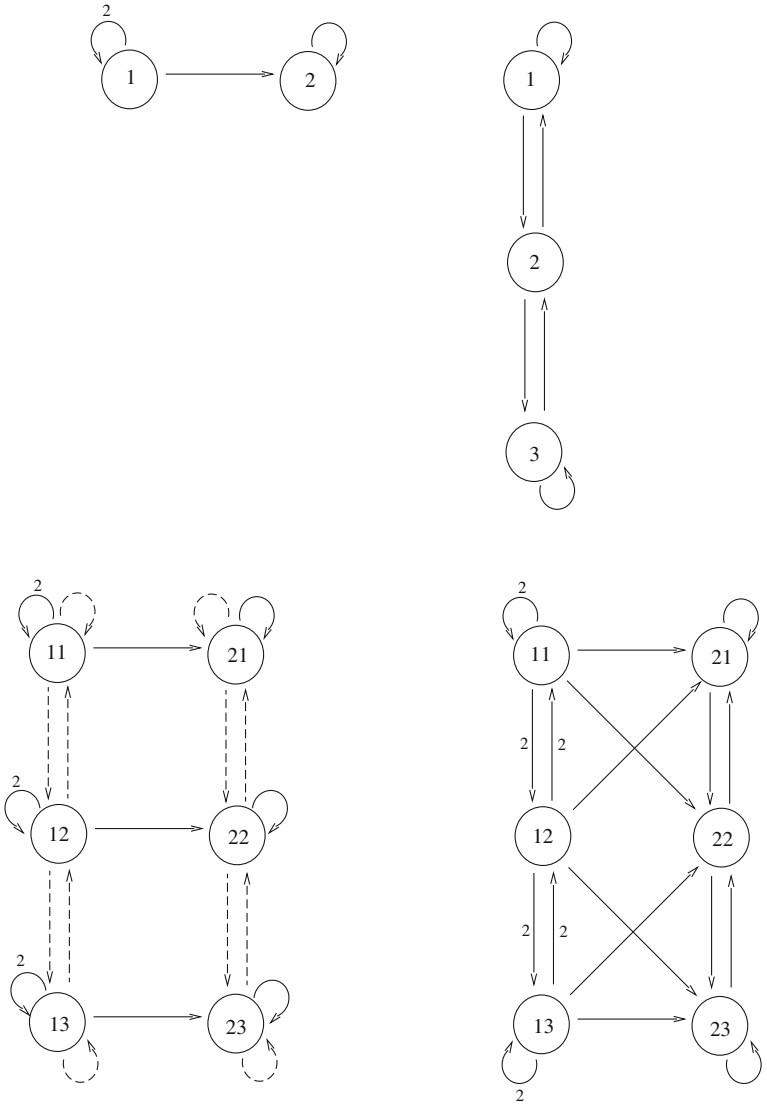


Fig. 7 From left to right: (up) networks N_1, N_2, (down) the cartesian product $N_1 \boxtimes N_2$ and the Kronecker product $N_1 \otimes N_2$

of the cartesian product presented in [3], since both edge types of R_1 and R_2 lead to just one edge-type in R. However, Theorem 6.5 in [3] that we apply next still holds in this case.) From [3, Theorem 6.5], the lattice of regular synchrony subspaces for R is given by the tensor product of the lattice of synchrony subspaces for R_1 and the lattice of synchrony subspaces for R_2. Given that the synchrony subspaces for

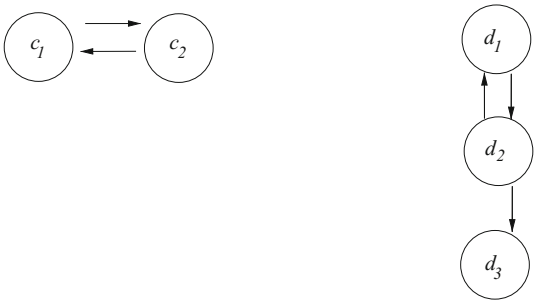


Fig. 8 Networks R_1 and R_2 such that the network R of Fig. 6 is the cartesian product of R_1 and R_2

R_1 are the trivial ones and that the two nontrivial synchrony subspaces for R_2 are defined by the coordinate equality condition $d_1 = d_2$ and $d_1 = d_3$, respectively, the nontrivial regular synchrony subspaces for R are thus the subspaces $\tilde{\Delta}_3$, $\tilde{\Delta}_8$, $\tilde{\Delta}_{12}$, $\tilde{\Delta}_{16}$, $\tilde{\Delta}_{19}$ and $\tilde{\Delta}_{23}$ from Table 2.

4.4 *f*-Join

The usual definition of *join of graphs* is given by the disjoint union of all graphs together with additional arrows added between every two cells from distinct graphs. In [7], Aguiar and Ruan introduce a generalized version of join on coupled cell networks.

Recall that a *multimap* is a generalized notion of map, where an element from the domain is assigned to a set of values from the range. Let $\tilde{C}_1 \subset C_1$ and $\tilde{C}_2 \subset C_2$ be non-empty subsets of cells. Denote by $P(\tilde{C}_2)$ the set of all subsets of $\tilde{C}_2$. Consider a multimap f from $\tilde{C}_1$ to $\tilde{C}_2$ given by

$$\begin{aligned} f &: \tilde{C}_1 \to P(\tilde{C}_2) \\ c &\mapsto f(c) \subset \tilde{C}_2. \end{aligned} \tag{3}$$

In [7], the *f-join* of two networks is defined as follows.

Definition 2 Let N_1 and N_2 be two identical-edge networks with set of cells C_1 and C_2, respectively, such that the cells in $C_1 \cup C_2$ are all of the same type and $C_1 \cap C_2 = \emptyset$. Let E_1 and E_2 be the set of edges of C_1 and C_2, respectively. A network N is called the *f-join* of N_1 and N_2, denoted by $N = N_1 *_f N_2$, if

- the set of cells of N is given by $C_1 \cup C_2$;
- the set of edges of N is given by $E_1 \cup E_2 \cup F$, where $F = \{(c, d), (d, c) : c \in \tilde{C}_1 \wedge d \in f(c)\}$ and f is defined by (3);
- if the edges in N_1 and N_2 are of the same type then any two edges e_1 and e_2 in N are of the same type; otherwise two edges e_1 and e_2 in N are of the same type if and only if they both are edges in E_1, in E_2 or in F.

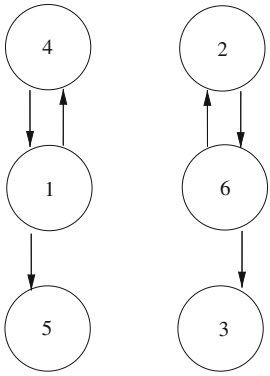


Fig. 9 The networks R_3 and R_4 such that the network R of Fig. 6 is the f-join of R_3 and R_4 where $f : \{4, 1, 5\} \to P(\{2, 6, 3\})$ with $f(4) = \{2\}$, $f(1) = \{6\}$, and $f(5) = \{3\}$

Note that, if $\tilde{C}_1 = C_1$, $\tilde{C}_2 = C_2$ and $f(c) \equiv C_2$ for all $c \in C_1$, then $N_1 *_f N_2$ is the join of N_1 and N_2, as defined for graphs.

Example 9 The network R of Fig. 6 may be seen as the f-join of two copies, R_3 and R_4, of the network R_2 on the right of Fig. 8, see Fig. 9. That is, $R = R_3 *_f R_4$ where $f : \{4, 1, 5\} \to P(\{2, 6, 3\})$ is the multimap such that $f(4) = \{2\}$, $f(1) = \{6\}$, and $f(5) = \{3\}$.

According to Definition 4.6 in [7], we can classify the synchrony subspaces of R into *non-bipartite*, *pairing bipartite* and *non-pairing bipartite*. A synchrony subspace is *non-bipartite* if, in its definition, there is no coordinate equality condition involving one cell in R_3 and one cell in R_4. A *bipartite* synchrony subspace is *pairing bipartite* if, in its definition, every coordinate equality condition involves one cell of R_3 and one cell of R_4 and for each cell there is at most one coordinate equality condition involving that cell, otherwise the synchrony subspace is said *non-pairing bipartite*.

The results in Theorem 4.17 of [7] characterize all the synchrony subspaces of $R = R_3 *_f R_4$. The non-bipartite and the pairing bipartite synchrony subspaces are easily obtained from these results, the synchrony subspaces of R_3 and R_4 and the interior symmetries of R.

The network R_3 has only two nontrivial synchrony subspaces, defined by the coordinate equality condition $x_1 = x_4$ and $x_4 = x_5$, respectively. Analogously, the network R_4 has only two nontrivial synchrony subspaces, defined by the coordinate equality condition $x_2 = x_6$ and $x_2 = x_3$, respectively.

Given a synchrony subspace S_1 for R_3 and a synchrony subspace S_2 for R_4, consider the polydiagonal of the total phase space of R defined by the conjunction of the coordinate equality conditions that define S_1 with the coordinate equality conditions that define S_2. From the results in Theorem 4.17 of [7], every non-bipartite synchrony subspace of R is such a polydiagonal subspace with the additional condition that if the coordinate equality conditions that define S_1 include $x_1 = x_4$ then the coordinate equality conditions that define S_2 must include $x_2 = x_6$ and if the coordinate equality conditions that define S_1 include $x_4 = x_5$ then the coordinate equality conditions

that define S_2 must include $x_2 = x_3$. We have then that the non-bipartite synchrony subspaces for R are $\tilde{\Delta}_6$, $\tilde{\Delta}_{19}$ and $\tilde{\Delta}_{23}$ from Table 2.

From the results in Theorem 4.17 of [7], every pairing bipartite synchrony subspace of R is given by some interior symmetry σ of R, where σ is a product of disjoint transpositions $\tau_i = (c_i, d_i)$ for $c_i \in \{4, 1, 5\}$, $d_i \in \{2, 6, 3\}$. There are five such interior symmetries of R: $\sigma_1 = (12)$, $\sigma_2 = (46)$, $\sigma_3 = (12)(46)$, $\sigma_4 = (16)(24)$, and $\sigma_5 = (16)(24)(35)$. We have then that the pairing bipartite synchrony subspaces for R are $\{\mathbf{x} : \ x_1 = x_2\}$, $\{\mathbf{x} : \ x_4 = x_6\}$ and $\{\mathbf{x} : \ x_1 = x_2, \ x_4 = x_6\}$ and the synchrony subspaces $\tilde{\Delta}_3$ and $\tilde{\Delta}_8$ from Table 2.

5 Complement Network

Suppose that N is a directed graph with n nodes and just with one edge-type, no multiarrows and no self loops. In graph theory, the usual definitions of the complement and converse graphs of G are the following:
(i) The *complement* of N is a directed graph $\overline{N}$ on the same set of nodes such that: a directed edge from node i to node j is present in $\overline{N}$ if it does not exist at N; a directed edge from node i to node j is not present in $\overline{N}$ if it exists at N. Graphically, if we take N and fill in all missing directed edges in order to obtain a complete graph (a simple directed graph in which every pair of distinct nodes is connected by a unique bidirectional edge), then $\overline{N}$ is obtained by removing the directed edges belonging to N. The sum of the $n \times n$ adjacency matrices of N and $\overline{N}$ is the $n \times n$ matrix with zero at the diagonal entries and 1 elsewhere and so it commutes with all $n \times n$ permutation matrices.
(ii) The *converse* of N is the graph with the same set of nodes as N and obtained from N by reversing the directions of all edges of N. The $n \times n$ adjacency matrix of the converse of N is the transpose of the adjacency matrix of N and so the sum of the two adjacency matrices, of N and its converse, is symmetric (it coincides with its transpose).

As an example of possible interpretation of the complement and converse graphs of a graph in the context of social networks is the following. The converse of a directed graph might be helpful in thinking about relations that have "opposites". The complement of a directed graph might be used to represent the absence of a tie. See for example Wasserman and Faust [45, p. 135].

Example 10 In Fig. 10 we show the complement (on the left) and the converse (on the right) of the five-cell network of Fig. 1.

Motivated by this, we define now the complement network of a network with n nodes that can have multiarrows, self-loops and more that one type of directed edges, preserving the fact that the sum of the adjacency matrices of the network and its complement, for each type of edges, is a matrix that commutes with all $n \times n$ permutation matrices – it can be seen as the adjacency matrix of an all-to-all

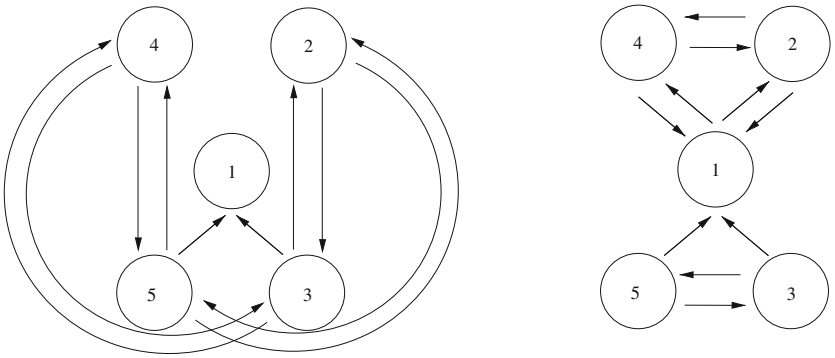


Fig. 10 The complement (on the left) and the converse (on the right) of the five-cell network of Fig. 1

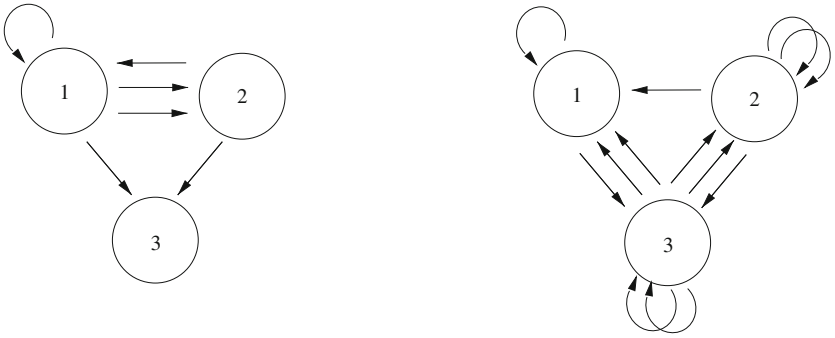


Fig. 11 A regular three-cell network with multiarrows and self loops at the left and its complement at the right

coupling n-cell network. In doing that, if the network corresponds to a directed graph just with one type of edges, no multiarrows and no self loops, then we recover the usual definition of the complement graph as just recalled above.

Definition 3 Let N be an identical-edge n-cell network with the set of cells $C = \{1, \ldots, n\}$ and adjacency matrix $M_N = [a_{ij}]$. Let $l = \max\{a_{ii} : i = 1, \ldots, n\}$ and $m = \max\{a_{ij} : i, j = 1, \ldots, n;\ i \neq j\}$. We define the *complement* network $\overline{N}$ to be the network with the set of cells C and the adjacency matrix $M_{\overline{N}}$ where $M_N + M_{\overline{N}}$ has at the diagonal entries $2l$ and m elsewhere. Graphically, if we take N and fill in all missing directed edges in order to obtain a graph where every cell has $2l$ self-loops and every two distinct cells have m bidirectional edges, then $\overline{N}$ is obtained by removing the directed edges belonging to G.

Example 11 In Fig. 11 we show a three-cell network with multiarrows and self loops at the left and its complement at the right.

As happens for the network N of Fig. 1 and its converse in Fig. 10, the converse of an homogeneous (regular) network may not be an homogeneous (regular) network. It follows, in particular, that, in general, a network N and its converse network do not have the same lattice of synchrony subspaces. Nevertheless, a network N and its converse have the same group of symmetries (but not necessarily the same group of interior symmetries).

For the complement network, we have the following:

Theorem 1 *Let N be an identical-edge network. Then, we have the following:*
(i) If N is regular then the complement network $\overline{N}$ is regular.
(ii) The networks N and $\overline{N}$ have the same lattice of synchrony subspaces.
(iii) The networks N and $\overline{N}$have the same group of symmetries.

Proof Let N be an identical-edge n-cell network with the set of cells $C = \{1, \ldots, n\}$ and adjacency matrix $M_N = [a_{ij}]$. Let $l = \max\{a_{ii} : \ i = 1, \ldots, n\}$ and $m = \max\{a_{ij} : \ i, j = 1, \ldots, n; \ i \neq j\}$. As before denote by $M_{\overline{N}}$ the adjacency matrix of its complement. By definition $M_N + M_{\overline{N}}$ has at the diagonal entries $2l$ and m elsewhere.
(i) Suppose N is regular of valency v. If $M_{\overline{N}} = [b_{ij}]$ then for $i, j = 1, \ldots, n$, we have: $b_{ii} = 2l - a_{ii}$ and if $i \neq j$ then $b_{ij} = m - a_{ij}$. It follows that for all i we have that $\sum_{j=1}^{n} b_{ij} = 2l - a_{ii} + \sum_{j \neq i}(m - a_{ij}) = 2l + (n-1)m - \sum_{j=1}^{n} a_{ij} = 2l + (n-1)m - v$. That is, the complement network $\overline{N}$ is regular of valency $2l + (n-1)m - v$.
(ii) A polydiagonal Δ in $\mathbf{R}^n$ represents a synchrony subspace of N (resp. $\overline{N}$) if and only if it is left invariant under M_N (resp. $M_{\overline{N}}$). Now the matrix $M_N + M_{\overline{N}}$ commutes with all $n \times n$ permutation matrices and so leaves invariant any polydiagonal. It follows then that a polydiagonal Δ is left invariant under M_N if and only if it is left invariant under $M_{\overline{N}}$. That is, Δ represents a synchrony space for N if and only if it represents a synchrony space for $\overline{N}$.
(iii) As the matrix $M_N + M_{\overline{N}}$ commutes with all $n \times n$ permutation matrices it follows then that a permutation matrix commutes with M_N if and only if it commutes with $M_{\overline{N}}$. □

We can generalize the above definition to homogeneous networks. Let N be an n-cell homogeneous network with k types of edges. Denote by $M_N^1, \ldots, M_N^k$ the k adjacency matrices of N, one for each edge type. It follows that N has k regular n-cell subnetworks, each with adjacency matrix M_N^i. Denote those by $N_1, \ldots, N_k$ and take $\overline{N}_1, \ldots, \overline{N}_k$ the corresponding complement networks. Then we can define the complement network $\overline{N}$ of N as the network with adjacency matrices $M_{\overline{N}_1}, \ldots, M_{\overline{N}_k}$.

Example 12 In Fig. 12 we show an homogeneous five-cell network at the left and its complement at the right.

Trivially, we have the following:

Corollary 1 *(i) The complement network of an homogeneous network is homogeneous.*

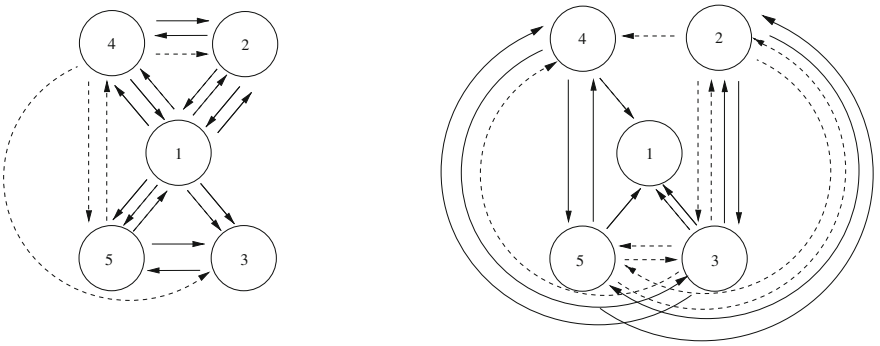


Fig. 12 An homogeneous five-cell network at the left and its complement at the right

(ii) An identical-cell network and its complement have the same lattice of synchrony subspaces.
(iii) An identical-cell network and its complement have the same group of symmetries.

Remark 1 (i) Note that, in case an n-cell network N is symmetric under a nontrivial finite group $\Gamma \subseteq \mathbf{S}_n$, the coupled cell systems associated with the network N and its complement are Γ-symmetric. It follows then that it can happen that the corresponding sets of dynamics supported by N and its complement are directly related, in the situations where the linear vector space of smooth Γ-symmetric vector fields coincide with both linear spaces of smooth vector fields with structure consistent with N and its complement, respectively.
(ii) Note that, in general, the fact that two coupled cell systems associated with N and $\overline{N}$, taking the same cell phase spaces, have the same set of synchrony subspaces, does not imply that the dynamics are closely related.

Acknowledgements The authors were partially funded 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 projects PTDC/MAT/100055/2008 and PEst-C/MAT/UI0144/2013.

References

1. Aguiar, M., Ashwin, P., Dias, A., Field, M.: Dynamics of coupled cell networks: synchrony, heteroclinic cycles and inflation. J. Nonlinear Sci. **21**(2), 271–323 (2011)
2. Aguiar, M.A.D., Dias, A.P.S.: The lattice of synchrony subspaces of a coupled cell network: characterization and computation algorithm. J. Nonlinear Sci. **24**(6), 949–996 (2014)
3. Aguiar, M.A.D., Dias, A.P.S.: Regular synchrony lattices for product coupled cell networks. Chaos **25**(1), 013108 (2015)
4. Aguiar, M.A.D., Dias, A.P.S., Golubitsky, M., Leite, M.C.A.: Homogeneous coupled cell networks with S3-symmetric quotient. Discrete Continuous Dynamical System. In: Proceedings

of the 6th AIMS International Conference Dynamical Systems and Differential Equations, suppl., pp. 1–9 (2007)
5. Aguiar, M.A.D., Dias, A.P.S., Golubitsky, M., Leite, M.C.A.: Bifurcations from regular quotient networks: a first insight. Phys. D **238**(2), 137–155 (2009)
6. Aguiar, M.A.D., Dias, A.P.S., Ruan, H.: Synchrony and elementary operations on coupled cell networks. SIAM J. Appl. Dyn. Syst. **15**(1), 322–337 (2016)
7. Aguiar, M.A.D., Ruan, H.: Evolution of synchrony under combination of coupled cell networks. Nonlinearity **25**(11), 3155–3187 (2012)
8. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
9. Aldis, J.W.: On balance. Ph.D. Thesis, University of Warwick (2009)
10. Antoneli, F., Dias, A.P.S., Paiva, R.: C: Hopf bifurcation in coupled cell networks with interior symmetries. SIAM J. Appl. Dyn. Syst. **7**(1), 220–248 (2008)
11. Antoneli, F., Dias, A. P. S., Paiva, R. C.: Coupled cell networks: Hopf bifurcation and interior symmetry. Discrete Continuous Dynamical System. In: 8th AIMS Conference Dynamical systems and Differential Equations and Applications, suppl., Vol. I, pp. 71–78 (2011)
12. Antoneli, F., Stewart, I.: Symmetry and synchrony in coupled cell networks. I. Fixed-point spaces. Int. J. Bifurc. Chaos Appl. Sci. Eng. **16**(3), 559–577 (2006)
13. Antoneli, F., Stewart, I.: Symmetry and synchrony in coupled cell networks. II. Group networks. Int. J. Bifurc. Chaos Appl. Sci. Eng. **17**(3), 935–951 (2007)
14. Antoneli, F., Stewart, I.: Symmetry and synchrony in coupled cell networks. III. Exotic patterns. Int. J. Bifurc. Chaos Appl. Sci. Eng. **18**(2), 363–373 (2008)
15. Arenas, A., Diaz-Guilera, A., Kurths, J., Moreno, Y., Zhou, C.: Synchronization in complex networks. Phys. Rep. **469**(3), 93–153 (2008)
16. Atay, F.M., Biyikoğlu, T.: Graph operations and synchronization of complex networks. Phys. Rev. E **(3) 72**(1), 016217 (2005)
17. Buono, P.L., Golubitsky, M.: Models of central pattern generators for quadruped locomotion: I. primary gaits. J. Math. Biol. **42**(4), 291–326 (2001)
18. Chen, G., Duan, Z.: Network synchronizability analysis: a graph-theoretic approach. Chaos **18**(3), 037102 (2008)
19. Dias, A.P.S., Lamb, J.S.W.: Local bifurcation in symmetric coupled cell networks: linear theory. Phys. D **223**(1), 93–108 (2006)
20. Dias, A.P.S., Moreira, C.S.: Spectrum of the elimination of loops and multiple arrows in coupled cell networks. Nonlinearity **25**(11), 3139–3154 (2012)
21. Dias, A.P.S., Paiva, R.C.: Hopf bifurcation in coupled cell networks with abelian symmetry. Bol. Soc. Port. Mat. Special Issue, 110–115 (2010)
22. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks. From Biological Nets to the Internet and WWW. Oxford University Press, Oxford (2003)
23. Field, M.: Combinatorial dynamics. Dyn. Syst. **19**(3), 217–243 (2004)
24. Golubitsky, M., Lauterbach, R.: Bifurcations from synchrony in homogeneous networks: linear theory. SIAM J. Appl. Dyn. Syst. **8**(1), 40–75 (2009)
25. Golubitsky, M., Nicol, M., Stewart, I.: Some curious phenomena in coupled cell systems. J. Nonlinear Sci. **14**(2), 207–236 (2004)
26. Golubitsky, M., Pivato, M., Stewart, I.: Interior symmetry and local bifurcation in coupled cell networks. Dyn. Syst. **19**(4), 389–407 (2004)
27. Golubitsky, M., Stewart, I.: The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space. Birkhauser, Basel (2002)
28. Golubitsky, M., Stewart, I.: Nonlinear dynamics of networks: the groupoid formalism. Bull. Am. Math. Soc. **43**(3), 305–364 (2006)
29. Golubitsky, G., Stewart, I., Buono, P.L., Collins, J.J.: Symmetry in locomotor central pattern generators and animal gaits. Nature **401**, 693–695 (1999)
30. Golubitsky, M., Stewart, I., Schaeffer, D.G.: Singularities and Groups in Bifurcation Theory. Vol. II. Applied Mathematical Sciences, vol. 69. Springer, New York (1988)

31. Golubitsky, M., Stewart, I., Török, A.: Patterns of synchrony in coupled cell networks with multiple arrows. SIAM J. Appl. Dyn. Syst. **4**(1), 78–100 (2005)
32. Hagberg, H., Schult, D.A.: Rewiring networks for synchronization. Chaos **18**(3), 037105 (2008)
33. Kamei, H.: Construction of lattices of balanced equivalence relations for regular homogeneous networks using lattice generators and lattices indices. Int. J. Bifurc. Chaos Appl. Sci. Eng. **19**(11), 3691–3705 (2009)
34. Kamei, H., Cock, P.J.A.: Computation of balanced equivalence relations and their lattice for a coupled cell network. SIAM J. Appl. Dyn. Syst. **12**(1), 352–382 (2013)
35. Leite, M.C.A., Golubitsky, M.: Homogeneous three-cell networks. Nonlinearity **19**(10), 2313–2363 (2006)
36. Lu, W., Atay, F.M., Jost, J.: Synchronization of discrete-time dynamical networks with time-varying couplings. SIAM J. Math. Anal. **39**(4), 1231–1259 (2007)
37. Moreira, C.S.: On bifurcations in lifts of regular uniform coupled cell networks. Proc. R. Soc. A **470**, 20140241 (2014)
38. Moreira, C.S.: Special jordan subspaces and synchrony subspaces in coupled cell networks. SIAM J. Appl. Dyn. Syst. **14**(1), 253–285 (2015)
39. Paiva, R.C.: Hopf bifurcation in coupled cell networks. Ph.D Thesis, University of Porto (2009)
40. Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization: A Universal Concept in Nonlinear Sciences. Cambridge Nonlinear Science Series, vol. 12. Cambridge University Press, Cambridge (2001)
41. Pogromsky, A., Santoboni, G., Nijmeijer, H.: Partial synchronization: from symmetry towards stability. Phys. D **172**(1–4), 65–87 (2002)
42. Restrepo, J.G., Ott, E., Hunt, B.R.: Emergence of synchronization in complex networks of interacting dynamical systems. Phys. D **224**(1–2), 114–122 (2006)
43. Stewart, I.: The lattice of balanced equivalence relations of a coupled cell network. Math. Proc. Camb. Philos. Soc. **143**(1), 165–183 (2007)
44. Stewart, I., Golubitsky, M., Pivato, M.: Symmetry groupoids and patterns of synchrony in coupled cell networks. SIAM J. Appl. Dyn. Syst. **2**, 609–646 (2003)
45. Wasserman, S., Faust, K.: Social Network Analysis Methods and Applications. Cambridge University Press, Cambridge (1994)

Inexact Subspace Iteration for the Consecutive Solution of Linear Systems with Changing Right-Hand Sides

Carlos Balsa, Michel Daydé, José M. L. M. Palma and Daniel Ruiz

Abstract We propose a two-phase acceleration technique for the solution of Symmetric and Positive Definite linear systems with multiple right-hand sides. In the first phase we compute some partial spectral information related to the ill conditioned part of the given coefficient matrix and, in the second phase, we use this information to improve the convergence of the Conjugate Gradient algorithm. This approach is adequate for large scale problems, like the simulation of time dependent differential equations, where it is necessary to solve consecutively several linear systems with the same coefficient matrix (or with matrices that present very close spectral properties) but with changing right-hand sides. To compute the spectral information, in the first phase, we combine the block Conjugate Gradient algorithm with the Inexact Subspace Iteration to build a purely iterative algorithm, that we call BlockCGSI. We proceed to an inner-outer convergence analysis and we show that it is possible to determine when to stop the inner iteration in order to achieve the targeted invariance in the outer iteration. The spectral information is used in a second phase to remove the effect of the smallest eigenvalues in two different ways: either by building a Spectral Low Rank Update preconditioner, or by performing a deflation of the initial residual in order to remove part of the solution corresponding to the smallest eigenvalues.

Keywords Inexact inverse iteration · Subspace iteration · Block conjugate gradient · Chebyshev filtering polynomials · Spectral projector

C. Balsa (✉)
Instituto Politécnico de Bragança (IPB), Bragança, Portugal
e-mail: balsa@ipb.pt

M. Daydé · D. Ruiz
IRIT, Universitée de Toulouse, CNRS, INPT, Toulouse, France
e-mail: dayde@enseeiht.fr

D. Ruiz
e-mail: ruiz@enseeiht.fr

J. M. L. M. Palma
Faculdade de Engenharia da Universidade do Porto (FEUP), Porto, Portugal
e-mail: jpalma@fe.up.pt

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_3

1 Introduction

We are interested in the computation of a *near*-invariant subspace associated with the smallest eigenvalues of a given symmetric and positive definite matrix, with a type of inexact subspace iteration method that exploits only matrix vector products. To this end, we exploit an algorithm, called BlockCGSI, which combines the inverse subspace iteration (SI), see [1] for instance, with a stabilized version of the block Conjugate Gradient algorithm (blockCG) [2, 3] to solve iteratively the set of multiple linear systems in each inverse iteration. The implicit use of the inverse of the coefficient matrix by means of an iterative solution (inner iteration), is suitable for large scale problems, where traditionally the factorization of the matrix is difficult to achieve, or when the matrix itself is not explicitly available. However, it also introduces an error - when computing the approximate solutions - that may affect the linear convergence of the inverse iteration (the outer iteration). The difficulty is to find an appropriate stopping threshold for the iterative method (in the inner iteration) that enables a suitable convergence of the inverse iteration and, if possible, that minimizes the computational work.

The central part in this work is to propose a way to monitor effectively the convergence of this inner-outer type of iterative scheme. We therefore analyze, from a geometrical point of view, the convergence of the inverse subspace iteration combined with the blockCG inner solver, and we derive an expression that relates the two residual norms used in the two iteration levels. From this, we can extract a residual measure for the blockCG that is directly linked with the convergence of the outer process toward the desired eigenvectors, and propose a stopping threshold parameter that minimizes the total amount of computational work to achieve some targeted accuracy.

In Sect. 2, we recall some important properties of the subspace iteration and of the inexact inverse iteration that are the basic components of the BlockCGSI algorithm. We also introduce some algorithmic techniques to improve the method. In particular, we exploit Chebyshev polynomials as a spectral filtering tool when building the starting vectors, and we introduce the concept of *"sliding window"* as an algorithmic feature for the computation of a *near*-invariant subspace of any dimension. Section 3 is dedicated to the analysis of the convergence properties. In particular, we explain how one should monitor the convergence of the blockCG in conjunction with the global convergence of the inverse iteration. The study presented in Sect. 3 follows on from our initial work on the monitoring of BlockCGSI presented in [4]. We finish the analysis of the BlockCGSI algorithm in Sect. 4, with a review of its main algebraic operations and computational costs.

The combination of the inverse subspace iteration with the block Conjugate Gradient (BlockCGSI algorithm) was initially exploited in the experimental study by [5], and used to deflate the initial residual in consecutive runs of the CG algorithm. This is of particular interest in the simulation of time dependent partial differential

equations, where at each global iteration (or time step) there are several systems with the same spectral properties to be solved.

Following this work, and to illustrate the effectiveness of the monitoring that we propose, we devote the last two sections to some numerical simulations where we first perform some partial spectral decomposition, and exploit this information to improve the convergence in the iterative solution of the following linear systems. In the case of Symmetric Positive Definite (SPD) matrices, several solutions have already been proposed to improve the convergence of Conjugate Gradient (CG) algorithm (see [6], for instance). In Sect. 5, we highlight two of these techniques, namely the deflation of the initial residual and the Spectral Low Rank Update (SLRU) preconditioner [7]. Section 6 is then concerned with the numerical results illustrating the proposed monitoring strategy for the BlockCGSI Algorithm as well as the potential of the pre-computed spectral information to accelerate the convergence of the Conjugate Gradient. We finish in Sect. 7 with some concluding remarks.

2 The Subspace (Inverse) Iteration

Subspace inverse iteration, or simply subspace iteration, is a generalization of the inverse iteration, where a set of vectors is multiplied consecutively by the inverse o a matrix instead of just one vector.

Consider the symmetric and positive definite matrices A and the matrix of the wanted eigenvectors $U = [u_1, u_2, \ldots, u_s]$, where $Au_j = \lambda_j u_j$, $j = 1, \ldots, s$. Let Z be a matrix whose columns generate a subspace of dimension s. If we multiply Z successively by A^{-1} we will generate a sequence $\{A^{-k}Z : k = 0, 1, 2, \ldots\}$ that converges to U. Defining the error angle between the approximated subspace generated by the columns of the matrix $A^{-k}Z$ and some specific eigenvector u_j as

$$\varphi_j^{(k)} \equiv \angle(u_j, A^{-k}Z) \equiv \min \angle(u_j, q) \qquad \text{over} \qquad q \in A^{-k}Z, \tag{1}$$

it is proved in [1, p. 333] that, under certain assumptions, each eigenvector u_j, $j \leq s$, satisfies

$$\tan\left(\varphi_j^{(k)}\right) \leq \left(\frac{\lambda_j}{\lambda_{s+1}}\right)^k \tan\left(\varphi_j^{(0)}\right). \tag{2}$$

Normally, after one (or several) multiplication by the matrix A^{-1}, the column vectors from the resulting matrix $Q^{(k)} = A^{-k}Z$ are orthonormalized. This simplest version of the subspace iteration is also called orthogonal iteration.

In order to get an optimal approximation to each individual eigenvector u_j, using all the information in the basis Q, the orthogonal iteration is normally followed by the Ritz (or Rayleigh-Ritz) acceleration. The resulting Ritz values $diag(\Delta) = \delta_1, \ldots, \delta_s$ and Ritz vectors $V = [v_1, v_2, \ldots, v_s]$ are approximations to the eigenpairs in A corresponding to the eigenvalues in the range $]0, \lambda_s]$. In [1, p. 334], it is indirectly

proved that $v_j^{(k)}$ converges linearly to u_j by proving that $v_j^{(k)} \longrightarrow q_j^{(k)}$ at the same asymptotic rate λ_j/λ_{s+1} that $q_j^{(k)}$ converges to $v_j^{(k)}$.

By (2), we can see that the reduction of the error angle in the subspace iteration is proportional to λ_j/λ_{s+1} instead of λ_s/λ_{s+1} as it is the case in the inverse iteration. This convergence property shows that if λ_j is well separated from λ_{s+1}, we can have a good estimation of the eigenvalue u_j in a few number of inverse iterations. In some cases, it can be useful to increase the block size s just to benefit from a better gap between λ_j and λ_{s+1}. This technique is also denoted as the use of *"Guard Vectors"*, e.g. extra vectors that are incorporated just to increase the rate of convergence in (2).

Compared with other reliable Lanczos algorithms, the subspace iteration just needs to store the current set of s approximated eigenvectors (Ritz Vectors). The previous vectors of the Krylov sequence are discarded. This can be an important advantage if we have to work with a low memory storage and with slow convergence. The main difficulty in the subspace iteration is that we must set a priori the working block size s. The block size defines the dimension of the targeted invariant subspace and also, as we can verify by (2), the convergence rate. As we don't know the eigenvalue distribution, the chosen block size can lead to a very slow convergence or, if the gap between the s wanted eigenvalues and the others is large, to a fast convergence where only a few subspace iterations will be needed for convergence. In this work we will also suggest a dynamical way to set up the block size s without any information a priori about the spectrum of the matrix A.

In the subspace iteration the Ritz values δ_j converge faster to their limit (the eigenvalue λ_j) than the corresponding Ritz vectors v_j to the eigenvector u_j. This means that one can have a converged Ritz value even if the corresponding Ritz vector is far from the wanted eigenvector. We can determine how close a Ritz vector v_j is close from the corresponding eigenvector u_j through the following error angle measure given by [1]

$$|\sin \angle(v_j, u_j)| \leq \frac{||Av_j - \delta_j v_j||_2}{gap}, \tag{3}$$

where $gap = \min\{|\lambda_{j-1} - \lambda_j|, |\lambda_j - \lambda_{j+1}|\}$. In practice we can not use (3) to monitor the convergence because the gap is unknown, but when the Ritz values have converged, we can use the δ_j's to approximate the gap and thus obtain a computable estimate of the error angle. However, for clustered eigenvalues, the spectral residual can be a bad measure because v_j can approximate another eigenvector different from u_j, and the formula (3) will not be reliable (see [1]). The error measure (3) also indicate that the angle between a Ritz vector v_j and the corresponding eigenvalue u_j is directly proportional to the invariance measure $||Av_j - \delta_j v_j||_2$.

In each subspace iteration a linear system with s right-hand sides is solved either by factorizing the coefficient matriz A or with an iterative solver. If it is solved iteratively, the error introduced by the inexact solution of the linear system may affect the convergence rate of the subspace iteration given by (2). The difficulty is to find an appropriate stopping threshold for the iterative method that enables a suitable convergence of the inverse iteration and, if possible, that minimizes the

global computational work. In the next section we proceed to an overview of the main ideas about this problematic in the case of working with a single vector (block size reduced to one), in which case the process is called inexact inverse iteration.

2.1 The Inexact Inverse Iteration

In the inverse iteration (subspace iteration with block size $s = 1$), the system of linear equation is traditionally solved through the factorization of the matrix A. This can be expensive or impractical if A has large dimension. Alternatively, we can solve these systems by an iterative method. Iterative methods are attractive in large scale problems because they require modest memory storage and the coefficient matrix does not need to be known explicitly, but just the result of it multiplication with any vector.

The inexact inverse iteration (see for instance [8]) includes two levels of iterations. One is the outer iteration, and corresponds to the main loop of the inverse iteration. The other is the inner iteration and corresponds to the iterative solution of the linear system in each outer iteration. The convergence of inexact inverse iteration is not yet perfectly well understood in all details, but has nevertheless been analyzed in several recent contributions. In this Section, we present a short survey of the main ideas exposed in some of these papers.

In [8, 9], for instance, it is proved that inexact inverse iteration can converge linearly at the same rate as the exact case even if the system is not solved accurately, and it is given some practical ways to choose the inner stopping threshold.

More recently in [10], a general convergence analysis of the correlation between the convergence rate and the threshold parameter is shown. It is proved that with some specific error measure, that if the threshold parameter is larger or equal to the ratio between the two smallest eigenvalues, the inexact inverse iteration converges linearly with a convergence rate directly given by the threshold parameter.

In [11], it is proved that it is worth continuing the inner loop until the norm of the solution vector stagnates. The growth of this norm is indeed directly linked to the reduction the eigenvalue residual norm in the inverse iteration. Consequently, the authors suggest a stopping criterion for the inner iteration based on the observation of the stagnation of the norm of approximate solution. Following this recommendation it is highlighted in [12] that this strategy is quite sensitive to the choice of the tolerance that measures this stagnation, as opposed to a strategy based on the measure of the standard relative residual of the system.

Similarly, the combination of the Jacobi–Davidson method with the Conjugate Gradient method as the inner solver has also been studied in [13]. It is established, from an analytical point of view, a relation between the reduction of the inner residual norm and the convergence of the outer process that allows an optimal stopping criterion for the inner iteration.

In most of these inexact inverse iteration analysis, the monitoring of the spectral error, in the outer iteration, is performed through some type of invariance measure of the approximated eigenvector v_1, like for instance

$$||Av_1 - \delta_1 v_1||_2 \leq \varepsilon, \tag{4}$$

where δ_1 is an approximation to the eigenvalue corresponding to v_1, like for instance the corresponding Ritz value. By Eq. (3) we know that this measure enables to control indirectly the error angle between v_1 from the corresponding eigenvector u_1. Note also that if v_1 is not orthonormalized the error measure (4) must be divided by $||v_1||_2$.

2.2 The BlockCGSI Algorithm

In this section, we present and detail partly the BlockCGSI algorithm (Algorithm 1) used to compute an M-orthonormal basis, represented by matrix W, of a *near*-invariant subspace associated with the smallest eigenvalues in the preconditioned matrix $M^{-1}A$, where M and A are both symmetric and positive definite. If this eigenspace incorporates, for instance, all the eigenvalues of $M^{-1}A$ in the range $]0, \mu[$, we can expect, when using it later as a second level of preconditioning, that the condition number of the coefficient matrix will be reduced to about $\kappa = \lambda_{\max}/\mu$ (where $\lambda_{\max}$ is the largest eigenvalue in $M^{-1}A$). In Algorithm 1, $\lambda_{\max}$ and μ are considered as input parameters. However there is no specific need to know exactly the largest eigenvalue, and some upper bound on $\lambda_{\max}$ is sufficient, provided it gives some rough estimation of the actual 2-norm of $M^{-1}A$. In our experiments, we simply set $\lambda_{\max} = 1$.

Another input concerns the choice of the block size s that defines the dimension of the working subspace at each inverse iteration. In the basic version of the inverse subspace algorithm, this also sets the number of approximated eigenvalues and eigenvectors at the end. Finally, it also gives the number of right-hand sides and solution vectors of the multiple linear systems solved by the blockCG algorithm at each inverse iteration, and therefore the amount of memory required as working space.

As a starting point, the algorithm requires the generation of an M-orthonormal matrix W of dimension s; the closer are these column vectors to the targeted *near*-invariant subspace, the faster the convergence of the inverse iteration will be. The scope of steps 1 to 4, in Algorithm 1, is to generate an initial M-orthonormal set $V^{(0)}$ of s vectors with eigencomponents corresponding to eigenvalues in the range $[\mu_f, \lambda_{\max}]$ below some predetermined value $\xi \ll 1$ (denoted as the *filtering level*). This filtering technique is based on Chebyshev polynomials (step 3) and is detailed in Sect. 2.4.

As we have seen, the essence of the inverse subspace iteration is the orthogonal iteration. It consists in multiplying a set of vectors by $A^{-1}M$ and M-orthonormalizing it in turn. If $W^{(k-1)}$ (initially empty) contains the set of vectors that have already converged at inverse iteration $(k-1)$, the current subspace $Q^{(k)}$, in step iii, should converge gradually to a *near*-invariant subspace that is M-orthogonal to $W^{(k-1)}$. In step i, the multiplication by $A^{-1}M$ is performed implicitly through the iterative solution of the system $M^{-1}AZ^{(k)} = V^{(k-1)}$ via the blockCG solver. In order to reduce the computational costs, this system is solved with an accuracy determined by the residual threshold ε. The appropriate choice of ε is detailed in Sect. 3.3.

ALGORITHM 1: BLOCKCGSI WITH SLIDING WINDOW

Inputs: A, $M = R^T R \in \mathbb{R}^{n\times n}$, μ, $\lambda_{max} \in \mathbb{R}$, $s \in \mathbb{N}$
Output: a *near*-invariant subspace W associated with all eigenvalues of $M^{-1}A$ in the range $]0, \mu]$

Begin

Generate the initial subspace (with filtering)

1. $Z^{(0)}$ =RANDOM(n, s)
2. $V^{(0)} = Z^{(0)}\Gamma$ such that $V^{(0)^T} V^{(0)} = I_{s\times s}$
3. $Q^{(0)}$=Chebyshev-Filter$(V^{(0)}, \xi, [\mu_f, \lambda_{\max}], A, R)$
4. $V^{(0)} = R^{-1} Q^{(0)}\Gamma$ such that $V^{(0)^T} M V^{(0)} = I_{s\times s}$
5. $W^{(0)}$ = empty
6. **For** $k = 1, ...,$ **until converge Do:**

Orthogonal iteration

i. Solve $M^{-1}AZ^{(k)} = V^{(k-1)}$ with BlockCG
ii. $P^{(k)} = Z^{(k)} - W^{(k-1)} W^{(k-1)^T} M Z^{(k)}$
iii. $Q^{(k)}\Gamma_k = P^{(k)}$ such that $Q^{(k)^T} M Q^{(k)} = I_{s\times s}$
iv. $Q^{(k)} = [W^{(k-1)}\ Q^{(k)}]$

Ritz acceleration

v. $\beta_k = Q^{(k)^T} A Q^{(k)}$
vi. Diagonalize $\beta_k = U_k \Delta_k U_k^T$
where $U_k^T = U_k^{-1}$
and Δ_k =Diag$(\delta_1, ..., \delta_{p+s})$ (Ritz Values)
vii. $V^{(k)} = Q^{(k)} U_k$ (Ritz Vectors)

"Sliding window"

viii. $W^{(k)}$ =converged columns of $V^{(k)}$
ix. $V^{(k)}$ =non-converged columns of $V^{(k)}$
x. (n, p) =size$(W^{(k)})$
xi. Update the computational window $(V^{(k)})$
xii. (n, s) =size$(V^{(k)})$

7. **EndDo**

End

In step ii, the approximate solution vectors $Z^{(k)}$ are then projected onto the orthogonal complement of the converged vectors $W^{(k-1)}$, in order to remove the influence of eigencomponents associated with the already converged eigenvalues. The set of projected vectors $P^{(k)}$ is then M-orthonormalized (step iii), and gathered together with $W^{(k-1)}$ in the matrix $Q^{(k)}$.

The orthogonal iteration is followed by the Ritz acceleration (steps v to vii). The spectral information contained in $Q^{(k)}$ is thus redistributed in the column vectors of $V^{(k)}$ that will contain separately better approximations of each individual eigenvector in the targeted invariant subspace. Steps v and vi give the Ritz values $diag(\Delta) = \delta_1, \ldots, \delta_{p+s}$ ranged in increasing order, where p is the dimension of $W^{(k-1)}$, i.e. the number of converged vectors in the inverse iteration $(k-1)$, and s is the current block size. The Ritz values and Ritz vectors $V = [v_1, v_2, \ldots, v_p, \ldots, v_{p+s}]$ are approximations to the eigenpairs in $M^{-1}A$ corresponding to the eigenvalues in the range $]0, \lambda_{p+s}]$. The convergence rate of each individual non-converged Ritz vector is then of order $\lambda_i/\lambda_{p+s+1}$, with $p+1 \leq i \leq p+s$.

The end of the BlockCGSI algorithm consists in testing the convergence and updating the computational window. In step viii, all the Ritz vectors that are considered as *near*-invariant, with respect to the given accuracy, are assigned to $W^{(k)}$. More details about the monitoring of the convergence are given in Sect. 3.1. Step xi consists in the update of the current set of vectors $V^{(k)}$. This algorithmic issue in the BlockCGSI algorithm is denoted as *"sliding window"* and detailed in Sect. 2.5.

2.3 Improvements of the BlockCGSI Algorithm

In this section, we describe briefly the two techniques incorporated in the BlockCGSI algorithm to improve the convergence: the Chebyshev based filtering technique at step 3 and the *"sliding window"* at step xi in Algorithm 1.

2.4 Chebyshev Based Filtering Technique

The purpose of the Chebyshev based filtering technique is to bring the randomly generated set of s starting vectors closer to the eigenvectors corresponding to s smallest eigenvalues. Chebyshev polynomials in $M^{-1}A$ are used to *damp* the eigenfrequencies associated with all the eigenvalues in the range $[\mu_f, \lambda_{max}]$, in the sense that those eigencomponents associated to all eigenvalues in this range are reduced to about 0, and the other ones are left close to their original value. We summarize here the outline of this technique, and for details, we refer to [6].

In the BlockCGSI algorithm, the application of the Chebyshev polynomial in $M^{-1}A$ to the set of starting vectors V is denoted as

$$Q = \text{Chebyshev-Filter}(V, \xi, [\mu_f, \lambda_{\max}], A, R),$$

with $M = R^T R$. This step can also be expressed formally by

$$Q = \mathscr{F}_m(M^{-1}A)V,$$

where $\mathscr{F}_m$ is a polynomial function of degree m given by

$$\mathscr{F}_m(\lambda) = \frac{T_m(w(\lambda))}{T_m(w(0))},$$

with T_m the usual Chebyshev polynomial of degree m and $w(\lambda)$ the mapping function that brings μ_f to 1 and $\lambda_{\max}$ to -1.

After the filtering process, the vectors $q_j = \mathscr{F}_m(M^{-1}A)v_j$ will have eigencomponents equal to $\mathscr{F}_m(\lambda_l)\zeta_j$, with $\zeta_j = \langle v_j, u_l \rangle, l = 1, \ldots, n$, and

$$\mathscr{F}_m(\lambda_l) \in \begin{cases} [-\xi, \xi] & \text{if } \lambda_l \in [\mu_f, \lambda_{\max}] \\ [\xi, 1] & \text{if } \lambda_l \in]0, \mu_f[, \end{cases}$$

where ξ, the filtering level, is chosen a priori much lower than 1 in order to make the eigencomponents corresponding to the eigenvalues in the range $[\mu_f, \lambda_{\max}]$ close to 0.

The Chebyshev polynomial T_m is computed implicitly by a recurrence formula from the two previous values T_{m-1} and T_{m-2} ($m \geq 2$). At each update, it requires that a set of s vectors be multiplied by $M^{-1}A$. For given values of μ_f, $\lambda_{\max}$ and ξ, the degree m depends on the ratio $\lambda_{\max}/\mu_f$ and is inversely proportional to ξ.

The reason behind the use of these Chebyshev filters at the starting point is to put the inverse subspace iteration in the situation of working directly in the orthogonal complement of a large number of eigenvectors, e.g. all those associated with the eigenvalues in the range $[\mu_f, \lambda_{\max}]$. Obviously, there is some compromise to achieve, in the sense that a very small value of μ_f will minimize the number of inverse iterations but will increase strongly the computational efforts in the Chebyshev filtering step.

2.5 Sliding Window

The original version of the BlockCGSI algorithm computes a fixed number s of approximated eigenvectors associated to the s smallest eigenvalues in the iteration matrix. The difficulty when choosing the parameter s is that we do not know a priori the distribution of the eigenvalues, and consequently how many eigenvalues we need approximate to reduce substantially the condition number. A too small block size s can lead to a non effective improvement in the convergence rate of the iterative solver in the following runs, whereas a too large block size s may induce unnecessary extra computational work. To circumvent this problem, we have included the possibility of enlarging the size of the *near*-invariant subspace along with the inverse iterations, as well as changing the block size s whenever appropriate. The idea is to start the algorithm with a block size s determined only on the basis of computer aspects like, for instance, the efficiency of Level-3 BLAS [14] internal kernels (used in the BlockCG algorithm), or the memory requirements. Then, when computing the Ritz values and checking the invariance of the Ritz vectors, at the end of each inverse iteration, we can decide how to adapt effectively the actual number of approximated eigenvectors.

In practice, when one or more of the s Ritz vectors in the current set $V^{(k)}$ are detected as *near*-invariant, these vectors are moved to the set of converged vectors $W^{(k)}$ (step viii of Algorithm 1), and there remains open the choice of incorporating new vectors to replace these ones to form the current block of s working vectors $V^{(k)}$, or to reduce the block size s (step xi of Algorithm 1). Incorporating new vectors is appropriate until the approximated eigenvalues cover a sufficiently large interval $[0, \mu]$ for an effective reduction of the condition number. When this target is met, it is then possible to reduce the block size s until all the targeted Ritz vectors have converged. However if we detect a gap in the actual range of the approximated eigenvalues, it can also be useful to keep the block size unchanged and to make use of the extra vectors to accelerate the convergence of the targeted ones in the inverse iteration. Effectively, with the *sliding window* the convergence rate given by (2) is now proportional to $\lambda_j/\lambda_{p+s+1}$, where p is the number of converged Ritz vectors.

Another issue in this algorithm comes from the fact that the solution of the linear systems in each inverse iteration are not obtained with high accuracy. Indeed, our purpose is to stop the blockCG as soon as possible. Consequently, it can happen that some of the Ritz values converge first to internal eigenvalues, before the smaller ones are actually discovered. In this case, some of the already converged Ritz vectors may appear as not enough invariant after the discovery of the extreme eigenvalues, and it may therefore be necessary to enlarge the block size s and refine furthermore these vectors. This risk of seeing internal eigenvalues coming first is the reason why it is important to systematically incorporate the assumed converged vectors $W^{(k-1)}$ when recomputing the Ritz pairs (step iv of Algorithm 1). This is the only way to ensure the appropriate redistribution of the eigencomponents within each approximate eigenvector in the long run.

The update of the computational window is mentioned at step xi of algorithm 1. The operation that consists in introducing new vectors, after a set of ℓ Ritz vectors has converged, is detailed in Algorithm 2.

ALGORITHM 2: INCORPORATE NEW VECTORS

Inputs: A, $M = R^T R \in \mathbb{R}^{n\times n}$, $V^{(k)} \in \mathbb{R}^{n\times(s-\ell)}\mu_f$, $\lambda_{max} \in \mathbb{R}$, $\ell \in \mathbb{N}$

Begin

a) P =RANDOM(n, ℓ)

b) $P = Q\Gamma$ such that $Q^T Q = I_{\ell\times\ell}$

c) P = Chebyshev-Filter(Q, ξ, $[\mu_f, \lambda_{\max}]$, A, R)

d) $Q = R^{-1} P\Gamma$ such that $Q^T M Q = I_{\ell\times\ell}$

e) $P = Q - W^{(k)} W^{(k)^T} M Q$

f) $V^{(k)} = [V^{(k)} P]$

End

It begins by generating randomly the new vectors and filtering them, as in the starting steps of the BlockCGSI algorithm (steps a, b and c). After that the vectors are projected in the M-orthogonal complement of the converged ones (step d and e), in order to remove the correspondent eigencomponents. The remaining operations (step f) adjust the block size s with respect to the current set of working vectors $V^{(k)}$.

3 Convergence Analysis

The BlockCGSI algorithm involves two iterative loops: the first, that we also denote as the outer iteration, at step 6 corresponds to the inverse iteration, and the second loop, or inner iteration, is in the call to the blockCG algorithm (at step i in Algorithm 1) for the iterative solution of the linear system with multiple right-hand sides, $M^{-1}AZ^{(k)} = V^{(k-1)}$. These two iterations levels require each some specific stopping criterion in order to monitor the convergence of the algorithm. Sections 3.1 and 3.2 are devoted to and analyze some properties associated with these aspects. In Sect. 3.3, we propose a way to link the monitoring of the convergence in the inner loop with the measure of the convergence in the outer loop.

3.1 Subspace Inverse Iteration (Outer Loop)

At each inverse iteration (k) in Algorithm 1, the blockCG algorithm solves the s linear systems $M^{-1}Az_j^{(k)} = v_j^{(k-1)}$, $j = 1, \ldots, s$, where the matrix A is preconditioned with a symmetric and positive definite preconditioner, $M = R^T R$. The symmetrized system can be written as usual as

$$R^{-T}AR^{-1}Rz_j^{(k)} = Rv_j^{(k-1)} \Longleftrightarrow \tilde{A}\tilde{z}_j^{(k)} = \tilde{v}_j^{(k-1)}, \qquad j = 1, \ldots, s. \tag{5}$$

where $\tilde{A} = R^{-T}AR^{-1}R, \tilde{z}_j = Rz_j$ and $\tilde{v}_j = Rv_j$. For simplicity we will omit to repeat that j varies from 1 to s. We will consider that the subscript j refers to the position of the correspondent eigenvalue in the current working set. The superscript (k) denotes the inverse iteration number, and the tilde refers to the symmetrized system (5).

The outer iteration produces a sequence of Ritz vectors $\tilde{v}_j^{(1)}, \tilde{v}_j^{(2)}, \ldots, \tilde{v}_j^{(k)}$, that converge to the eigenvector $\tilde{u}_j = Ru_j$ corresponding to the eigenvalue λ_j of both matrices $\tilde{A}$ and A. At the outer iteration (k), the vectors $\tilde{v}_j^{(k)}$ are orthonormal, while the vectors $v_j^{(k)}$ (columns of matrix $V^{(k)}$, in step vii of Algorithm 1) are M-orthonormal. The corresponding Ritz values (diagonal elements of the matrix $\Delta^{(k)}$) are given by $\delta_j{}^{(k)} = v_j{}^{(k)^T} Av_j^{(k)} = \tilde{v}_j^{(k)T}\tilde{A}\tilde{v}_j^{(k)}$.

As we have seen in Sect. 2, we can evaluate indirectly the error angle (between v_j and the corresponding eigenvector u_j) through the measure

$$\frac{||\tilde{A}\tilde{v}_j^{(k)} - \delta_j^{(k)}\tilde{v}_j^{(k)}||_2}{||\tilde{v}_j^{(k)}||_2} = ||M^{-1}Av_j^{(k)} - \delta_j^{(k)}v_j^{(k)}||_M. \tag{6}$$

Dividing (6) by $\delta_j^{(k)}$ (as an approximation of λ_j) we obtain a relative invariance measure that is used in step viii of Algorithm 1 to decide if a Ritz vector $v_j^{(k)}$ has converged or not, as for instance when

$$\frac{||M^{-1}Av_j^{(k)} - \delta_j^{(k)} v_j^{(k)}||_M}{\delta_j^{(k)}} \leq \varepsilon_{\text{outer}}, \tag{7}$$

if a certain tolerance $\varepsilon_{\text{outer}}$ is enough. Since $\varepsilon_{\text{outer}}$ is fixed, the error angle will be smaller for the Ritz vectors corresponding to the smallest eigenvalues because, in these cases, the invariance measure (6) will be divided by a smaller values δ_j. We use the stopping criterion (7) because, as we will see in Sect. 5, the Ritz vectors are used to improve the convergence of the CG algorithm, and this measure of *near*-invariance in the Ritz vectors is indeed very appropriate in that respect. Note also that (7) is also used in the same context by [11].

3.2 The BlockCG Iteration (Inner Loop)

The block Conjugate Gradient (blockCG) algorithm under concern is a numerically stable variant [3] that avoids the numerical problems that can occur when some of the s systems are about to converge. It solves simultaneously the s linear systems from Eq. (5). For each system, $j = 1, .., s$, it produces a sequence of vectors $\tilde{z}_j^{[i]}$, giving, after convergence, the approximate solution $\tilde{z}_j^{(k)}$ used in the kth inverse iteration,

$$\tilde{z}_j^{[1]}, \tilde{z}_j^{[2]}, \ldots, \tilde{z}_j^{[i]} \rightarrow \tilde{z}_j^{(k)}, \tag{8}$$

where the superscript $[i]$ stands for the blockCG (inner) iteration number.

The residual vector associated with each iterate $\tilde{z}_j^{[i]}$ is

$$\tilde{r}_j^{[i]} = \tilde{v}_j^{(k-1)} - \tilde{A}\tilde{z}_j^{[i]}. \tag{9}$$

We also introduce another vector which we will use to measure the proximity of the current iterate $\tilde{z}_j^{[i]}$ from the corresponding eigenvector $\tilde{u}_j$,

$$\tilde{S}_j^{[i]} = \tilde{A}\tilde{z}_j^{[i]} - \tilde{\delta}_j^{[i]}\tilde{z}_j^{[i]} = \tilde{v}_j^{(k-1)} - \tilde{\delta}_j^{[i]}\tilde{z}_j^{[i]} - \tilde{r}_j^{[i]}, \tag{10}$$

where $\tilde{\delta}_j^{[i]} = \tilde{z}_j^{[i]T}\tilde{A}\tilde{z}_j^{[i]} / \tilde{z}_j^{[i]T}\tilde{z}_j^{[i]} = \delta_j^{[i]}$ is the Rayleigh quotient corresponding to the current iterate $\tilde{z}_j^{[i]}$. The error measure (6) applied on the symmetrized system, at each inner iteration $[i]$, yields

$$\frac{||\tilde{S}_j^{[i]}||_2}{||\tilde{z}_j^{[i]}||_2} = \frac{||\tilde{v_j}^{(k-1)} - \delta_j^{[i]}\tilde{z}_j^{[i]} - \tilde{r}_j^{[i]}||_2}{||\tilde{z}_j^{[i]}||_2}. \tag{11}$$

Additionally, if we start the blockCG iteration with $\tilde{z}_j^{[0]} = 0$, at each iteration the current residual $\tilde{r}_j^{[i]}$ remains orthogonal to both $\tilde{v_j}^{(k-1)}$ (the right-hand side) and $\tilde{z}_j^{[i]}$ (linear

combination of the current Krylov vectors). Thus, $\tilde{r}_j^{[i]T}(\tilde{v_j}^{(k-1)} - \delta_j^{[i]}\tilde{z}_j^{[i]}) = 0$, and we can write

$$||\tilde{v_j}^{(k-1)} - \delta_j^{[i]}\tilde{z}_j^{[i]} - \tilde{r}_j^{[i]}||_2^2 = ||\tilde{v_j}^{(k-1)} - \delta_j^{[i]}\tilde{z}_j^{[i]}||_2^2 + ||\tilde{r}_j^{[i]}||_2^2. \tag{12}$$

Finally, if we translate the previous properties to the non-symmetrized system,

$$\frac{||M^{-1}Az_j^{[i]} - \delta_j^{[i]}z_j^{[i]}||_M}{||z_j^{[i]}||_M} = \sqrt{\frac{||v_j^{(k-1)} - \delta_j^{[i]}z_j^{[i]}||_M^2}{||z_j^{[i]}||_M^2} + \frac{||r_j^{[i]}||_M^2}{||z_j^{[i]}||_M^2}} \stackrel{def}{=} \sqrt{\phi_j^{[i]2} + \omega_j^{[i]2}}, \tag{13}$$

where we can see that the reduction of the invariance of each iterate $z_j^{[i]}$ during the inner loop depends on the relative residual measure $\omega_j^{[i]} = ||r_j^{[i]}||_M/||z_j^{[i]}||_M$ and on the value $\phi_j^{[i]} = ||v_j^{(k-1)} - \delta_j^{[i]}z_j^{[i]}||_M/||z_j^{[i]}||_M$.

Even if we expect that the backward error measure $\omega_j^{[i]}$ will decrease down to a level of small magnitude, the value of $\phi_j^{[i]}$ is more likely to stagnate on a higher level, depending on the proximity of the right-hand side $v_j^{(k-1)}$ from the correspondent eigenvector u_j. Therefore, the bound in (13) can be dominated by the value of $\phi_j^{[i]}$, and little improvement on the global convergence of the algorithm can be expected by further iterations in the blockCG.

We now investigate the asymptotic behavior of $\phi_j^{[i]}$, assuming that $z_j^{[i]}$ actually converges to $z_j^* = A^{-1}Mv_j^{(k-1)}$. Let us first introduce the asymptotic limit of the Rayleigh quotient $\delta_j^{[i]}$, $\delta_j^* = \langle z_j^*, v_j^{(k-1)}\rangle_M/||z_j^*||_M^2$, and the angle θ_j in the M-norm between z_j^* and $v_j^{(k-1)}$, whose cosine is given by

$$\cos(\theta_j) = \frac{\langle z_j^*, v_j^{(k-1)}\rangle_M}{||z_j^*||_M||v_j^{(k-1)}||_M} = \delta_j^*||z_j^*||_M. \tag{14}$$

As a consequence of the M-orthonormalization of $v_j^{(k-1)}$, we can write

$$\sin(\theta_j) = ||v_j^{(k-1)} - \delta_j^* z_j^*||_M, \tag{15}$$

which is also the asymptotic limit of $||v_j^{(k-1)} - \delta_j^{[i]}z_j^{[i]}||_M$. Consequently, the asymptotic limit of the component $\phi_j^{[i]}$ in (13) is

$$\phi_j^{[i]} \xrightarrow[i\to\infty]{} \frac{\sin(\theta_j)}{||z_j^*||_M} = \delta_j^* \tan(\theta_j). \tag{16}$$

We can see that the asymptotic limit of $\phi_j^{[i]}$ depends only on the two vectors $v_j^{(k-1)}$ and $z_j^* = A^{-1}Mv_j^{(k-1)}$. It is also clear that, if $v_j^{(k-1)}$ is close to an eigenvector, the angle θ_j should be very small, as well as the corresponding asymptotic limit of $\phi_j^{[i]}$. With respect to the bound in (13), this allows more room for decreasing the backward error $\omega_j^{[i]}$ in the blockCG iteration. The strategy suggested by this analysis is to decrease the value of the stopping criterion in the blockCG (inner loop) along with the convergence of the inverse iteration (outer loop). This basic idea is further developed in the next section.

3.3 Stopping Criterion for the BlockCG

The stopping criterion for the blockCG defines the approximation degree of $\tilde{z}^{(k)} \approx \tilde{A}^{-1}\tilde{v}_j^{(k-1)}$ or equivalently of $z^{(k)} \approx A^{-1}Mv_j^{(k-1)}$. Its choice is crucial because demanding a high accuracy can lead to a great amount of unnecessary extra work, whereas an insufficient level of accuracy in the solution may deteriorate the convergence rate of the inverse iteration (given by (2)).

We propose to monitor only the convergence of the iterates $z_1^{[i]}$, corresponding to the smallest Ritz value $\delta_1^{(k-1)}$ in the previous inverse iteration. In general, this system needs more computational efforts to be solved accurately. As we have seen in the previous section, the relative residual in the inner loop is given by

$$\omega_1^{[i]} = \frac{||v_1^{(k-1)} - M^{-1}Az_1^{[i]}||_M}{||z_1^{[i]}||_M}, \tag{17}$$

and is readily available in the blockCG iteration (see [3]). Notice also that $\omega_1^{[i]}$ is very close to the usual Rigal–Gaches [15] backward error measure

$$\frac{||v_1^{(k-1)} - M^{-1}Az_1^{[i]}||_M}{||z_1^{[i]}||_M + 1},$$

using the fact that the M-norm of the current right-hand side $v_1^{(k-1)}$ equals 1, and assuming that the 2-norm of the preconditioned matrix $M^{-1}A$ is close to one.

In the outer loop, we monitor the accuracy of the approximated eigenvectors through the relative invariance, as indicated in (6) and (7). At inverse iteration $(k-1)$, we consider that a Ritz vector $v_j^{(k-1)}$ has converged when

$$\frac{||M^{-1}Av_j^{(k-1)} - \delta_j^{(k-1)}v_j^{(k-1)}||_M}{\delta_j^{(k-1)}} \leq \varepsilon_{\text{outer}}. \tag{18}$$

In order to satisfy (18) in the current inverse iteration (k), the stopping criterion in the blockCG is set as

$$\omega_1^{[i]} \leq \varepsilon_{\text{inner}}, \quad \text{with} \quad \varepsilon_{\text{inner}} = \varepsilon_{\text{outer}} \delta_1^{(k-1)}, \tag{19}$$

where $\delta_1^{(k-1)}$ is the smallest of the Ritz values corresponding to current set of non converged Ritz vectors. This stopping criterion is based on the decomposition (13), assuming that $\phi_1^{[i]}$ is not dominant and that the value of $w_1^{[i]}$ governs the absolute invariance measure in the inner iteration. This is surely the case when the Ritz vector $v_1^{(k-1)}$ is close to an eigenvector because, in this case, the value of $\tan(\theta_1)$ should be small. This can even occur at the first inverse iteration because the starting vectors are previously filtered with Chebyshev based polynomials. A second assumption, implicit in (19), is that the inner invariance measure, given by Eq. (13), will be close to the outer invariance measure, given by Eq. (6). This can be justified from the theoretical analysis given in [1], which shows that the Ritz vectors $V^{(k)}$ actually converge to the solution vectors $Z^{(k)}$ at the same rate of convergence of these $Z^{(k)}$ vectors to the set of targeted eigenvectors (see Sect. 2).

Under these assumptions, the idea in (19) is to achieve a given tolerance $\varepsilon_{\text{outer}}$ in (18) while minimizing the number of blockCG iterations. Note also that the strategy (19) for the stopping criterion is in agreement with other analysis of the inexact inverse iteration, like for instance in [8, 10, 12], where it is suggested to use a decreasing value for the inner tolerance $\varepsilon_{\text{inner}}$. This is the case, indeed, in (19) since the value of $\delta_1^{(k)}$ decreases gradually toward λ_1 along with the convergence of the outer iteration.

As we have seen in Sect. 3.2, when $\phi_1^{[i]}$ has reached its stagnation level defined in (16), the bound in (13) is then dominated by this asymptotic value, and there is no need to decrease any further the value of $\omega_1^{[i]}$. No more improvements on $v_1^{(k)}$ with respect to u_1 might be expected, and it is better to stop the blockCG iteration and to launch the next inverse iteration. This strategy is very close to the one proposed in [13] and, in some way, to the one also proposed in [11], because the stagnation of $\phi_1^{[i]}$ implies the stagnation of $||z_1^{[i]}||_M$ which is the basic argument used in this article to monitor the convergence.

The risk of having $\omega_1^{[i]}$ much smaller than $\phi_1^{[i]}$ during the blockCG iterations is also limited with a closer tolerance $\varepsilon_{\text{outer}}$ not too small, like for instance 10^{-1} or 10^{-2}. These values are in general enough for the purpose of building a *near*-invariant subspace for preconditioning the solution of consecutive linear systems with the same coefficient matrix. However, if one is interested in computing an accurate invariant subspace, the inner threshold parameter $\varepsilon_{\text{inner}}$ in (19) should be set to the maximum between $\varepsilon_{\text{outer}} \delta_1^{(k-1)}$ and the asymptotic value of $\phi_1^{[i]}$ in (16). From the analysis in 3.2, this is indeed the maximum level of accuracy that it is reasonable to achieve in each blockCG run. Do not forget also that along with the convergence of the Ritz vectors towards the corresponding eigenvectors, the asymptotic value of $\phi_1^{[i]}$ in (16) tends to zero proportionally to the tangent of the angle θ_j. Because of that, it is ensured that after some appropriate number of inverse (outer) iteration, the maximum between $\varepsilon_{\text{outer}} \delta_1^{(k-1)}$ and this asymptotic value of $\phi_1^{[i]}$ will always remain the first of these two, and the targeted accuracy in the inverse iteration can then be expected to be achieved afterward.

Table 1 Basic operations counts in BlockCGSI algorithm

Operation	Size	Flops	Symbol	BLAS level
$x = \text{RANDOM}$	n	$3n$	$\mathscr{C}_{RAND}$	–
$y \leftarrow y^T x$	n	$2n$	$\mathscr{C}_{DOT}$	1
$y \leftarrow y + \sigma x$	n	$2n$	$\mathscr{C}_{AXPY}$	1
$y = Ax$	n	$2nnz(A) - n$	$\mathscr{C}_{A}$	2
$x = M^{-1} y$	n	$4nnz(R) - 2n$	$\mathscr{C}_{M}$	2
$C \leftarrow C + \sigma VB$	$n \times s$	$2s^2 n$	$\mathscr{C}_{GEMM}$	3
$P = QR$	$n \times s$	$2s^2 n$	$\mathscr{C}_{ORTHO(s)}$	1

4 Operations Counts

The precomputation of the basis W of a *near*-invariant subspace with the BlockCGSI algorithm, has a cost which we denote by $\mathscr{C}_{BCGSI}$. For a fixed accuracy, this cost depends essentially on the dimension q of the basis W and on some working parameters like the block size s, the filtering level ξ and the cut-off filtering value μ_f. To be effective, the gains obtained in the acceleration of the convergence of the classical Conjugate Gradient algorithm must cover the extra cost for the computation of this spectral information. In Table 1 we present the costs in floating point operations (flops) associated with some basic operations performed in the BlockCGSI algorithm, as well as the corresponding BLAS level. The computational cost of each part in Algorithm 1 will be expressed as a function of these basic operations. We denote the computational cost of one operation OP by the symbol $\mathscr{C}_{OP}$, like for instance $\mathscr{C}_A$ that is the cost of on sparse matrix-vector product, where the number of non-zeros elements of A is given by $nnz(A)$. As mentioned before, a first level of left preconditioner M is also used, which purpose is to cluster the spectrum of our iteration matrix. The cost of the multiplication of a vector by M^{-1} is represented by $\mathscr{C}_M$. Since, M is constructed in our experiments by means of the Incomplete Cholesky factorization or Jacobi scaling, its cost can be estimated as mentioned in Table 1, where $nnz(R)$ is the number of nonzero elements in the factor R from the Incomplete Cholesky or Jacobi factorization.

In the beginning of the BlockCGSI algorithm, we apply the Chebyshev filtering polynomial in A to bring the set of s random generated vectors $V^{(0)}$ near the eigenvectors corresponding to the smallest eigenvalues. The cost of one Chebyshev iteration is $\mathscr{C}_{CHEBY} \approx s\mathscr{C}_A + s\mathscr{C}_M + 3s\mathscr{C}_{AXPY}$. Additionally, the starting vectors are orthonormalized before the filtering step and M-orthonormalized after. The computation of the starting vectors has a total cost given by

$$\mathscr{C}_{START} \approx s\mathscr{C}_{RAND} + 2\mathscr{C}_{ORTHO(s)} + \frac{s}{2}\mathscr{C}_M + \texttt{ChebIt} \times \mathscr{C}_{CHEBY}, \qquad (20)$$

where `ChebIt` is the total number of Chebyshev filtering iterations.

The QR iteration represents the most expensive step in Algorithm 1 in terms of computational cost. It consists of the iterative solution of the system with s right-hand sides using the stabilized blockCG solver. The cost of each inner iteration in the blockCG is:

$$\mathscr{C}_{bCG} \approx s\mathscr{C}_A + s\mathscr{C}_M + 3\mathscr{C}_{GEMM} + 2\mathscr{C}_{ORTHO(s)}. \tag{21}$$

After the blockCG run, the solution vectors $Z^{(k)}$ are projected onto the M-orthogonal complement of the converged Ritz vectors $W^{(k-1)}$ (of dimension p, that varies from 0 to $q-1$), and are M-orthonormalized. The estimation of the operations count corresponding to the steps included in the QR iteration at each inverse iteration resumes in:

$$\mathscr{C}_{QR} \approx \texttt{bCGIt(k)} \times \mathscr{C}_{bCG} + s\mathscr{C}_{PROJ(p)} + \mathscr{C}_{ORTHO(s)} + \mathscr{C}_M, \tag{22}$$

where `bCGIt(k)` is the number of blockCG inner iterations performed at inverse iteration (k), and where $\mathscr{C}_{PROJ(p)} = \mathscr{C}_M + p\mathscr{C}_{DOT} + \mathscr{C}_{AXPY}$. The cost of the QR iteration cannot be determined a priori, because the parameters `bCGIt(k)` and p change from one inverse iteration to the other in a non-deterministic way. Therefore, we trace their actual values as the algorithm progresses, and we compute the total number of flops at the end. We proceed in the same way in the Ritz acceleration with the size of $Q^{(k)}$, given by $s+p$. The cost of one Ritz acceleration is given by

$$\mathscr{C}_{RITZ} \approx (s+p)\mathscr{C}_A + 2(s+p)^2\mathscr{C}_{DOT} + 5(s+p)^3, \tag{23}$$

where $5(s+p)^3$ is the cost corresponding to the spectral decomposition of the matrix β_k at step vi in Algorithm 1.

The amount of work to update the computational window is induced by the convergence test, Eqs. (6) and (7), and by the incorporation of the new vectors (see Algorithm 2). If we maintain the same block size s, the ℓ converged vectors are replaced in the computational window by ℓ new vectors. After these new vectors are filtered and M-orthonormalized, as done with the initial set of starting vectors, they are finally projected onto the M-orthogonal complement of the converged ones. The cost of these steps is then given by

$$\mathscr{C}_{UPDATE} \approx \mathscr{C}_{INV} + \ell\mathscr{C}_{RAND} + \texttt{ChebIt} \times \mathscr{C}_{CHEBY} + 2\mathscr{C}_{ORTHO(\ell)} + \frac{\ell}{2}\mathscr{C}_M + \ell\mathscr{C}_{PROJ(p)}, \tag{24}$$

where

$$\mathscr{C}_{INV} \approx 2(s+p)\mathscr{C}_M + (s+p)\mathscr{C}_{AXPY} + 2(s+p)\mathscr{C}_{DOT} \tag{25}$$

is the cost spent when testing the invariance of all the Ritz vectors.

Finally, the estimate of the total number of floating point operations performed in the BlockCGSI algorithm, over all the inverse iterations `InvIt`, is

$$\mathscr{C}_{BCGSI} \approx \mathscr{C}_{START} + \texttt{InvIt} \times (\mathscr{C}_{QR} + \mathscr{C}_{RITZ} + \mathscr{C}_{UPDATE}). \tag{26}$$

5 Exploiting the Spectral Information

Once the *near*-invariant subspace linked to the smallest eigenvalues of the linear system is obtained, we can use it for solving any system with the same coefficient matrix, taking advantage of this spectral information to remove the effect of the poor conditioning. An overview of techniques that exploit this idea to improve the convergence of the Conjugate Gradient can be found, for instance, in [6]. Here, we summarize two of these techniques based on the same approach, i.e. building a spectral projector that enables to work in the orthogonal complement of the invariant subspace corresponding to the smallest eigenvalues.

5.1 Deflated Starting Guess

One of the possible methods is to compute a starting guess, by means of an oblique projection of the initial residual ($r^{(0)} = M^{-1}b - M^{-1}Ax^{(0)}$) onto the *near*-invariant subspace associated with the eigenvalues in the range $]0, \mu[$, to get the corresponding eigencomponents in the system solution:

$$\begin{aligned} r^{(1)} &= r^{(0)} - M^{-1}AW\Delta^{-1}W^T Mr^{(0)}, \\ \text{and } x^{(1)} &= x^{(0)} + W\Delta^{-1}W^T Mr^{(0)}, \end{aligned}$$

where W and Δ are the matrices of the q converged Ritz vectors and values, obtained by the BlockCGSI algorithm on the preconditioned matrix $M^{-1}A$. To compute the remaining part $x^{(2)}$ of the exact solution vector $x^* = x^{(1)} + x^{(2)}$, we can solve $M^{-1}Ax^{(2)} = r^{(1)}$ with the Conjugate Gradient algorithm.

In practice, as $x^{(0)} = 0$ we run the Conjugate Gradient to solve $M^{-1}Ax = M^{-1}b$ starting from the deflated component of the solution

$$x^{(1)} = W\Delta^{-1}W^T b. \tag{27}$$

With this initial starting guess, we expect that the CG will converge to the remaining part of the solution very quickly, since the difficulties caused by the smallest eigenvalues have been swallowed, and the eigenvalues bounds are given by μ and $\lambda_{\max}$, as explained in [16]. In this way, the Conjugate Gradient should reach linear convergence immediately [17]. For clarity, we will call INIT- CG the algorithm corresponding to CG with a starting guess obtained with deflation.

5.2 Spectral Low Rank Update (SLRU) Preconditioner

Another way of exploiting spectral information from the coefficient matrix is to perform a deflation at each CG iteration, instead of just at the beginning. This approach, proposed in [7], is called Spectral Low Rank Update (SLRU) preconditioning.

The computation of the solution of the preconditioned system $M^{-1}Ax = M^{-1}b$ is obtained by means of the CG algorithm applied to an equivalent system $\widehat{M}Ax = \widehat{M}b$, where the preconditioner $\widehat{M}$ is given by

$$\widehat{M} = M^{-1} + W\Delta^{-1}W^T. \tag{28}$$

In this case M and $\widehat{M}$ are also called the first and second level of preconditioning. The preconditioner $\widehat{M}$ will shift the smallest eigenvalues in the coefficient matrix $M^{-1}A$ close to one (see [7]). In some cases, it can be useful to shift the smallest eigenvalues close to some predetermined value λ (with $\lambda = \lambda_{max}$, for instance), in which case the spectral preconditioner should be set to

$$\widehat{M} = M^{-1} + \lambda W\Delta^{-1}W^T.$$

This is not useful in our test problem, because the spectrum is previously clustered near one with the first level of preconditioning M. In the following, we will call SLRU- CG to the algorithm corresponding to the preconditioned Conjugate Gradient with SLRU as preconditioner.

5.3 Pratical Considerations

We first consider operations count. As in Sect. 4, we assume that $q \ll n$ so that we neglect terms not containing n, (q is the dimension of the *near*-invariant basis W).

In addition to the sparse matrix-vector product with A at each iteration, the CG iteration add merely two dot-products, DOT, and three vector updates, AXPY (see Sect. 4).

The algorithms INIT- CG and SLRU- CG perform both an oblique projection of the initial residual onto the *near*-invariant basis W of size q, involving the pre-computation of $W\Delta^{-1}W^Tb$, where $\Delta = W^TAW$ is the Ritz matrix computed on step vi of Algorithm 1. The cost of its inversion is not significant because we consider it as a diagonal matrix. We note that Δ^{-1} is stored jointly with the basis W. In the scheme SLRU-CG, the multiplication with the preconditioner $\widehat{M}$ also implies an oblique projection $W\Delta^{-1}W^Tr^{(k)}$ at each iteration (see Eq. (28)). This projection can be done using common level 2 BLAS operations [18] with a total cost roughly equal to

$$\mathscr{C}_{\text{Proj}} \approx 4\,(p+1)\,n. \tag{29}$$

In Table 2, we indicate the total cost in floating-point operations for each algorithm, with an initial cost, and a cost per iteration. One of the major differences between INIT-CG and SLRU- CG is that SLRU- CG uses the projection operator $W\Delta^{-1}W^T$ at each iteration, whereas INIT- CG does exploit this only at the beginning (for computing a starting vector $x^{(1)}$, see formula (27)). Anyway, this also contributes to better numerical stability in the convergence process of SLRU- CG.

Table 2 Cost in floating-point operations for different methods

	The cost in floating-point operations	
	At the beginning	At each iteration
CG	$\mathscr{C}_A + \mathscr{C}_M + - + -$	$\mathscr{C}_A + \mathscr{C}_M + 3\mathscr{C}_{AXPY} + 2\mathscr{C}_{DOT} + -$
INIT- CG	$\mathscr{C}_A + \mathscr{C}_M + \mathscr{C}_{\text{Proj}} + -$	$\mathscr{C}_A + \mathscr{C}_M + 3\mathscr{C}_{AXPY} + 2\mathscr{C}_{DOT} + -$
SLRU- CG	$\mathscr{C}_A + \mathscr{C}_M + \mathscr{C}_{\text{Proj}} + -$	$\mathscr{C}_A + \mathscr{C}_M + 3\mathscr{C}_{AXPY} + 2\mathscr{C}_{DOT} + \mathscr{C}_{\text{Proj}}$

Let us briefly examine the memory requirements of these two acceleration techniques. In comparison with the CG algorithm, the INIT- CG and SLRU- CG schemes require about the same amount of extra storage, of order $n(q+1)$, to store W, the basis of the *near*-invariant subspace (Ritz vectors), and $diag(\Delta)$ the corresponding Ritz Vectors.

6 Numerical Experiments

In this section, we report on some numerical experiments concerning the computation of the spectral information associated with the smallest eigenvalues of a preconditioned matrix $M^{-1}A$. We also include some experiments concerning the use of the pre-computed spectral information to improve the consecutive solution of linear systems with the same coefficient matrix as mentioned in Sect. 5. Theses aspects are illustrated on a test matrix coming from the 2D heterogeneous diffusion equation in a L shape region, discretized by finite elements, with size $n = 7969$. Another application of the BlockCGSI algorithm to a larger problem can be found in [19].

We also precondition the resulting linear system with Jacobi scaling or classical Incomplete Cholesky (see Table 3).

We divide the experiments in two parts, the first one concerns the monitoring of the BlockCGSI algorithm discussed in Sect. 3 and the second concerns the improvement of the convergence of the CG algorithm with the pre-computed spectral information.

Table 3 Properties of the test matrix $M^{-1}A$ with two different preconditioners

Preconditioner	$\lambda_{\min}$	$\lambda_{\max}$	# eigs. below $\mu =$			nnz	nnz	n
			$1e-3$	$5e-3$	$1e-2$	R	A	
Jacobi	$3.07e-09$	2.08	2	9	18	7969	55131	7969
IC(0)	$1.66e-08$	1.55	2	2	3	31550		

6.1 Monitoring the Subspace Iteration

In Fig. 1, we show the convergence behavior of the inner invariance measure

$$\frac{||M^{-1}Az_1^{[i]} - \delta_1^{[i]} z_1^{[i]}||_M}{||z_1^{[i]}||_M}, \tag{30}$$

which is directly related with the values of ω_1 and ϕ_1 as evidenced in (13). The two plots in Fig. 1 illustrate the behavior of these three values in the blockCG run at the first inverse iteration ($k = 1$). The first plot corresponds to the case of non filtered starting vectors, and the second one to the case of starting vectors filtered with a level $\xi = 1e-2$ and a cut-off value $\mu_f = 5e-3$.

We can observe the effect of the Chebyshev filtering of the starting vectors, which helps to make the value of ϕ_1 much smaller than what it can be with a randomly generated initial set of vectors. The direct consequence is that ω_1 then becomes a good measure of the inner invariance measure (30), even at the very beginning of the algorithm. Additionally, the filtering of the starting vectors changes the convergence behavior of the blockCG, because the filtered right-hand sides have more favorable spectral properties. It also enables to decrease substantially the asymptotic value of ϕ_1 in the first inverse iteration, allowing a larger range of values for the choice of the threshold $\varepsilon_{\text{inner}}$ in the blockCG, which is a desirable feature for the algorithm as discussed in Sect. 3.3.

In Fig. 2 we plot the evolution of the invariance measure (6) of the Ritz vector v_1 as a function of the number of inverse iterations. We compare two different inner stopping criteria with the exact inverse iteration. The lozenge curve corresponds to the stopping criteria (19), the circles curve corresponds to stop the inner iteration when $\omega_1 < \phi_1$ and the circles curve corresponds to the exact case (performed by the Cholesky factorization of the coefficient matrix). We can observe in Fig. 2 that the proposed inner stopping

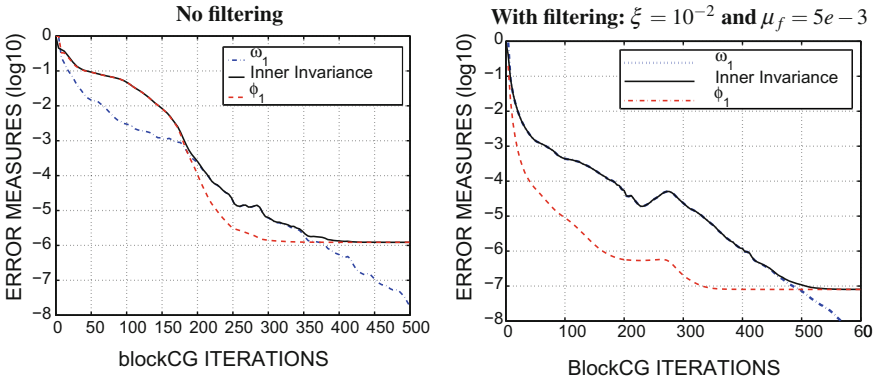


Fig. 1 Correlation between the inner invariance measure (30), ω_1 and ϕ_1, in the blockCG with block size 4, and at the first subspace iteration. The test matrix is preconditioned with Jacobi scaling

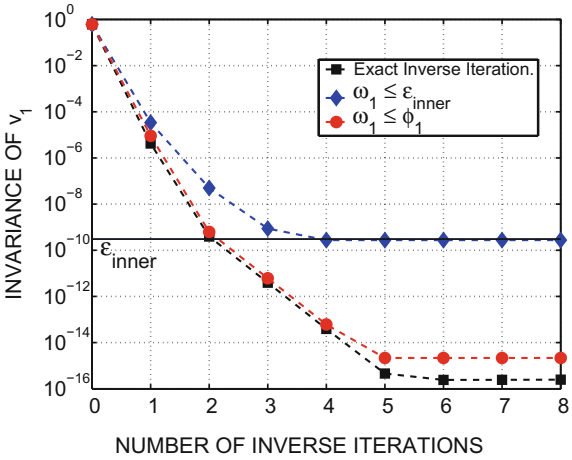


Fig. 2 Invariance of the Ritz vector v_1 over the number of outer iterations

criteria force the invariance to decrease until the targeted tolerance is reached, which occur at the 4th inverse iteration. Stopping the blockCG when $\omega_1 < \phi_1$ enables to reach the same behavior as in the exact case, which is in agreement with the analysis developed in Sect. 3.2.

In Table 4, we present both the total number of inner and outer iterations in the BlockCGSI algorithm (see Algorithm 1) to compute a *near*-invariant subspace associated with all eigenvalues in the range $]0, \mu[$. The requested relative invariance in these approximated eigenvectors was set to $\varepsilon_{\text{outer}} = 10^{-1}$, with respect to the convergence criterion for the outer loop given by Eq. (18), and the stopping criterion for the blockCG set accordingly as in (19). The total number of inverse iterations is indicated by `InvIt`, and the value of `bCGIt` denotes the sum of all the iterations performed by the blockCG solver in the given BlockCGSI run. The Chebyshev iterations count, `ChebIt`, incorporates all the Chebyshev iterations spent when filtering the starting vectors, as well as when incorporating new vectors during update of the computational window (see Algorithm 2). Finally, we also include the total number of floating point operations performed by the BlockCGSI algorithm ($\mathscr{C}_{BCGSI}$), in millions (`Mflops`), computed as in (26). We varied the filtering level from $\xi = 1e-6$ to $\xi = 1e-16$, including the case of no filtering. The block size was chosen to illustrate these cases, e.g. when it is below, equal or greater than the targeted number of eigenvectors (q). The two cut-off values of the filtering step μ_f correspond to the cases when it is greater or equal to the principal cut-off value μ in Algorithm 1.

The results in Table 4 show that the algorithm manages to compute the targeted spectral information independently of the choice of the block size s. Of course, it is optimal when s is correlated to the actual number of eigenvalues (q) in the range $]0, \mu[$. In this case, all the iterations counts are minimized as well as the total number of operations. With larger block sizes s, the algorithm benefits from the *"guard vectors"* effect (see Sect. 2), and the number of inverse iterations are reduced. A greater block size also improves the convergence of the block Conjugate Gradient. For these reasons, the

Table 4 Iteration and operation counts as a function of the filtering level ξ

Filter level ξ	$\mu = 1.0e-3$ (2 eigenvalues)							
	$s = 2, \mu_f = 1.0e-2$				$s = 5, \mu_f = 1.0e-2$			
	InvIt	bCGIt	ChebIt	Mflops	InvIt	bCGIt	ChebIt	Mflops
–	6	240	–	138	2	177	–	469
$1e-6$	2	170	105	131	2	98	105	347
$1e-8$	2	139	138	125	2	67	138	294
$1e-10$	2	109	171	119	2	50	171	278
$1e-12$	2	80	205	114	2	26	205	245
$1e-14$	2	53	238	110	2	10	238	230
$1e-16$	2	35	275	120	2	9	275	256
Filter level ξ	$\mu = 5.0e-2$ (9 eigenvalues)							
	$s = 5, \mu_f = 1.0e-2$				$s = 9, \mu_f = 1.0e-2$			
	InvIt	bCGIt	ChebIt	Mflops	InvIt	bCGIt	ChebIt	Mflops
–	29	432	–	1347	14	247	–	2033
$1e-6$	10	119	630	1333	2	65	105	661
$1e-8$	4	69	414	777	2	41	138	530
$1e-10$	3	51	342	572	2	23	171	444
$1e-12$	3	27	410	595	2	6	205	367
$1e-14$	3	11	476	636	2	3	238	387
$1e-16$	3	10	542	715	1	3	271	436
Matrix preconditioned with Jacobi scaling								
Filter level ξ	$\mu = 1.0e-2$ (3 eigenvalues)							
	$s = 3, \mu_f = 1.0e-1$				$s = 5, \mu_f = 1.0e-1$			
	InvIt	bCGIt	ChebIt	Mflops	InvIt	bCGIt	ChebIt	Mflops
–	6	96	–	148	4	73	–	243
$1e-6$	6	75	28	143	2	44	28	180
$1e-8$	6	63	37	134	2	35	37	165
$1e-10$	6	53	46	127	2	27	46	153
$1e-12$	3	35	55	99	2	19	55	140
$1e-14$	3	27	64	96	2	12	64	128
$1e-16$	2	19	75	88	1	10	73	130
Filter level ξ	$\mu = 3.0e-2$ (9 eigenvalues)							
	$s = 5, \mu_f = \mu$				$s = 9, \mu_f = \mu$			
	InvIt	bCGIt	ChebIt	Mflops	InvIt	bCGIt	ChebIt	Mflops
–	30	191	–	907	17	118	–	1310
$1e-6$	4	41	104	358	2	25	52	351
$1e-8$	3	20	138	352	2	11	69	274
$1e-10$	3	12	170	390	1	3	85	235
$1e-12$	3	6	204	434	1	3	102	275
$1e-14$	2	4	236	450	1	1	118	295
$1e-16$	2	4	270	556	1	1	135	335
Matrix preconditioned with $IC(0)$								

increase of s does not necessarily imply an increase of the total amount of work. When the block size is smaller than q, the *"sliding window"* feature enables to obtain at any rate all the targeted vectors. Our experiments also show that the final number of converged vectors can exceed the actual number of eigenvalues in the range $]0, \mu[$ when the block size is not equal to this number.

As we can observe in Table 4, the filtering of the new vectors with Chebyshev polynomials can improve quite a lot the efficiency of the BlockCGSI algorithm. As the filtering level ξ decreases, the number of inverse iterations is reduced because the resulting filtered vectors get closer to a *near*-invariant subspace, and the stagnation level of ϕ_j becomes much lower (see also Fig. 1). This also gives room for larger decrease of the inner invariance (30) at each inverse iteration. When the block size s is equal to the number q of eigenvalues in $]0, \mu[$, the convergence can be reached with a minimum number of inverse iterations (`InvIt` $= 1$ for instance). The number of blockCG iterations is also reduced because of the better spectral properties of the right-hand sides. Obviously, decreasing the filtering level ξ also increases the number of Chebyshev iterations in the filtering process. In that respect, there is a compromise to reach in terms of total computational cost. The optimal value of ξ, that minimizes this computational work (`Mflops`), also depends on the other filtering parameter μ_f, and on the number of targeted eigenvalues q. We have observed that it is better in general to take the value of the initial filtering parameter μ_f not too large with respect to the actual bound μ on the range of targeted eigenvalues, and also that when the number of targeted eigenvalues is small, a good filtering level ξ is indicated.

6.2 Improving the CG Convergence

Based on the pre-computed spectral information, we can improve the convergence of the CG algorithm. To illustrate this, we have solved the two tests systems with both INIT- CG or SLRU- CG. In Fig. 3, we plot the backward error

$$\rho^{(i)} = \frac{||M^{-1}r^{(i)}||_M}{||M^{-1}r^{(0)}||_M} = \frac{||R^{-T}r^{(i)}||_2}{||R^{-T}r^{(0)}||_2}, \tag{31}$$

normally used in the preconditioned Conjugate Gradient, where $r^{(i)} = b - Ax^{(i)}$.

The results show the effectiveness of the use of the spectral information to reduce the total number of iterations. When the system is preconditioned with Jacobi scaling (see plots on the top of Fig. 3), the initial condition number, of order 10^8, can even be reduced to the order of $10^3 \approx \lambda_{max}/\mu$ with the use of the first two vectors associated with the two smallest eigenvalues only. With the use of a larger number of Ritz vectors, the reduced condition number is maintained to about the same level, and just little improvements can be expected in the convergence rate of the CG algorithm. With the Incomplete Cholesky preconditioner, the number of critical eigenvalues seems also to be 2 (see plots on the bottom of Fig. 3).

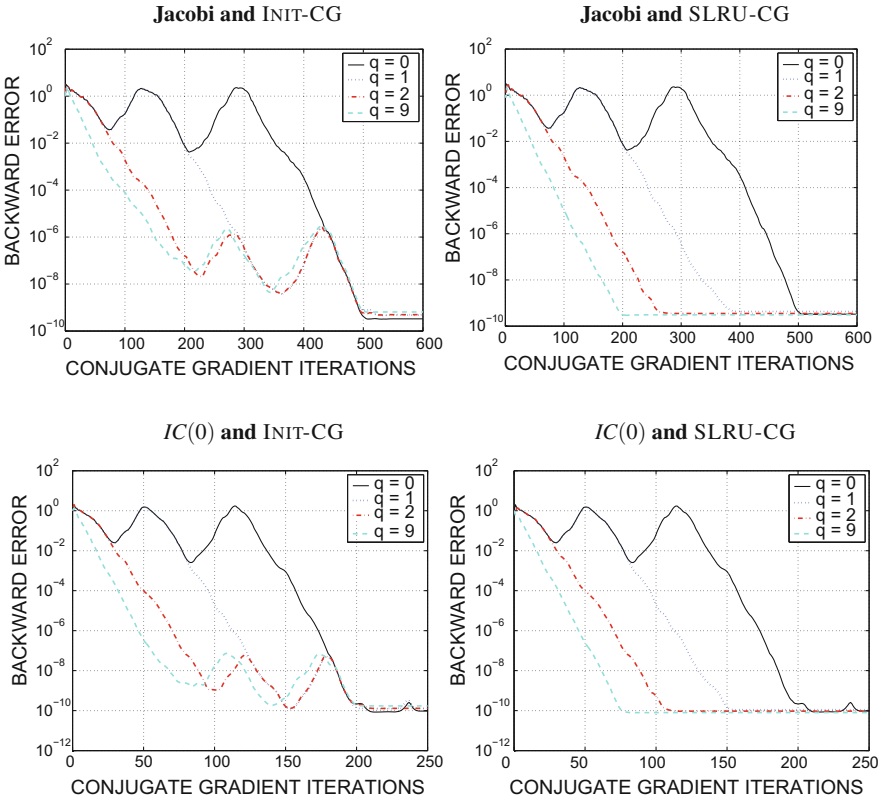


Fig. 3 Convergence behavior of INIT- CG and SLRU- CG with different sizes q of the pre-computed *near*-invariant subspace. The system is initially preconditioned either with Jacobi scaling (top) or with the standard Incomplete Cholesky (bottom)

Regarding the results in Fig. 3, the SLRU preconditioner is numerically more stable than the deflation technique. The SLRU- CG converges linearly while the INIT- CG looses the linear rate of convergence when reaching small residual values (say 10^{-6} with Jacobi preconditioner and 10^{-8} with $IC(0)$). To maintain this linear rate all through the iterations, the spectral information needs to be more accurate, as we can observe in Fig. 4. Indeed, with larger values for the number of correct digits t (see formula (18)) the irregularities in the rate of convergence of INIT- CG are smoothed gradually. However the cost for computing a much more accurate *near*-invariant basis can be rather large, and it can be preferable to simply use the SLRU- CG algorithm if solutions with high precision are needed.

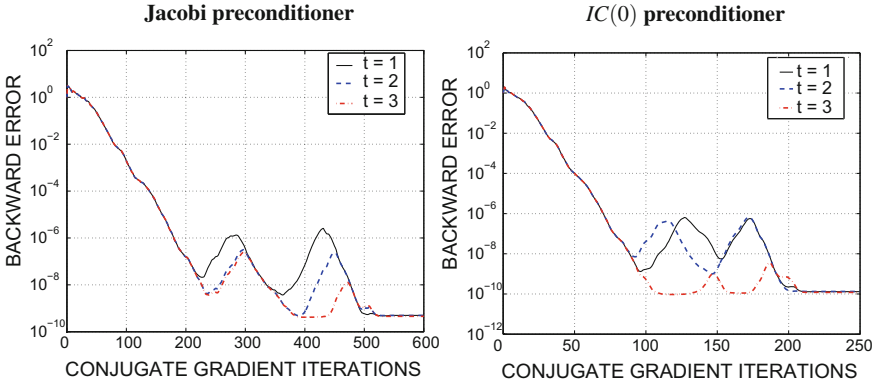


Fig. 4 Effect of the accuracy of the spectral information on the convergence behavior of INIT- CG

6.3 Cost-Benefit

We have proposed a technique to improve the consecutive solutions of several systems with the same coefficient matrix but with different right-hand sides. This technique is based on a two phase approach: we first perform a partial spectral decomposition of the coefficient matrix $M^{-1}A$ with the help of BlockCGSI algorithm, and we use this information afterward to accelerate the CG through the deflation of the starting guess or with a second level of preconditioning. We illustrate how the gains obtained at each solve can reduce substantially the total computational cost in the long run.

We begin by presenting the costs in floating-point operations involved in each CG run (see Sect. 5.3). In Fig. 5, we plot the history of the backward error versus the number of floating-point operations. In the case of SLRU- CG algorithm, we observe that higher dimension q of the *near*-invariant subspace does not always bring an improvement in the convergence. The oblique projection (28) performed at each iteration is responsible for the growth of the computational work when the dimension q gets larger. As we can see on the right of Fig. 5, when q varies from 3 to 20 the rate of convergence decreases, and no gains are obtained despite the effective reduction of the number of CG iterations (see Fig. 3). As opposed to that, we can observe in the case of INIT- CG algorithm (left of Fig. 5) that we always get improvements with larger values of q.

As we have seen in Sect. 4, the pre-computation of the *near*-invariant subspace W has a cost, that we denote by $\mathscr{C}_{BCGSI}$, depends on the dimension of this subspace and on some working parameters in the BlockCGSI algorithm. To be effective, the gains obtained in the acceleration of the convergence of the given iterative solvers must compensate, in some way, the extra cost for this spectral pre-computation. In Table 5, we present the computational costs $\mathscr{C}_{BCGSI}$ (in millions of operations, `Mflops`) for three different cases that correspond to different choices for the cut-off value μ. For each one, we have computed all the q Ritz vectors corresponding to the q eigenvalues in the range $]0, \mu[$. The spectral information is computed with two correct digits ($t = 2$) in the case of Jacobi preconditioner, and with one digit ($t = 1$) in the case of $IC(0)$.

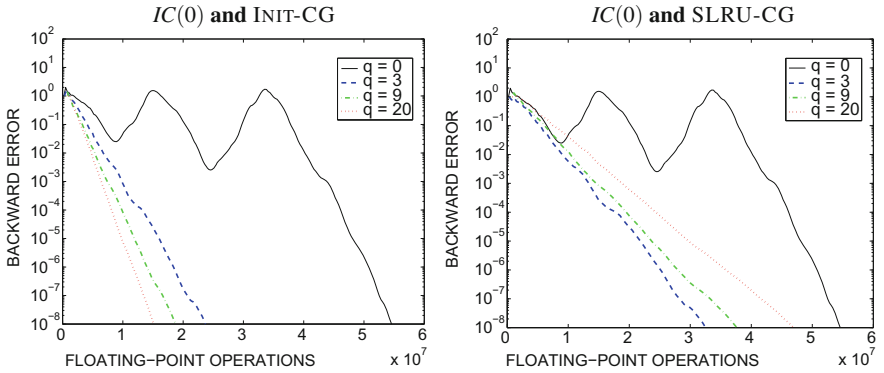


Fig. 5 History of the backward error as a function of the computational cost, for different sizes q of the *near*-invariant subspace. The case $q = 0$ corresponds to the CG algorithm, for comparison

To illustrate the cost-benefits, we stop the CG iterations when the relative residual norm $\rho^{(i)}$ (see Eq. (31)) is below 10^{-8}. When preconditioned with Jacobi scaling, the Conjugate Gradient performs 478 iterations with a cost of 95 `Mflops`, and when preconditioned with $IC(0)$, 187 iterations with a cost of 55 `Mflops`. In this table, we indicate the number of CG iterations (`Nit`), the number of floating-point operations `Mflops` and the number of amortization right-hand sides (`Amor. rhs.`), i.e., the number of right-hand sides that have to be considered in consecutive solves before the extra cost $\mathscr{C}_{BCGSI}$ is compensated. The number of amortization right-hand sides is given by

$$\texttt{Amor. rhs.} = \left\lfloor \frac{\mathscr{C}_{BCGSI}}{\mathscr{C}_{CG} - \mathscr{C}_{aCG}} \right\rfloor + 1,$$

where $\mathscr{C}_{CG}$ is the cost of CG algorithm without acceleration, and $\mathscr{C}_{aCG}$ is cost of the accelerated CG, either INIT- CG or SLRU- CG algorithm. These informations are given for each cut-off value μ. For instance, in Table 5, with Jacobi scaling and $\mu = 4.0e-3$, 211 `Mflops` are needed for the spectral pre-computation, out of which the INIT- CG algorithm achieves convergence in 190 iterations and 38 `Mflops`, i.e. a reduction of 60% compared to the run which does not use this spectral information. Consequently, the 211 extra `Mflops` are paid back after four consecutive accelerated solves, compared to four runs of the non-accelerated CG.

Table 5 shows that Init-CG and SLRU-CG converge, in general, in the same number of iterations if the spectral information enough accurate. The main difference is that SLRU demands more `Mflops` as the size of the *near*-invariant subspace (q) increases. For this reason, the number of amortization vectors (`Amor. rhs`) is greater with the SLRU preconditioner. In the case of the chosen stopping threshold (10^{-8}), the INIT- CG approach seems to be preferable.

In summary, the numerical results demonstrate that, when an accuracy of order 10^{-8} is required to solve linear systems in sequence with the same matrix but with changing right-hand sides, the cost of pre-computation of a *near*-invariant subspace with Block-

Table 5 Cost-benefits of CG accelerated with the spectral information

	Spectral fact.			Deflated $x^{(0)}$			SLRU prec.		
	μ	q	$\mathscr{C}_{BCGSI}$	CG		Amor. rhs.	CG		Amor. rhs.
			Mflops	Nit	Mflops		Nit	Mflops	
Jacobi precond.	–	0	–	478	95	–	478	95	–
	$1.0e-3$	2	184	227	45	4	227	63	6
	$4.0e-3$	5	211	190	38	4	187	71	9
	$5.5e-3$	9	427	176	38	8	166	84	39
$IC(0)$ precond.	–	0	–	187	55	–	187	55	–
	$1.0e-2$	3	88	89	26	4	80	33	4
	$2.0e-2$	5	145	79	23	5	74	35	8
	$3.0e-2$	9	235	75	22	8	62	37	14

CGSI algorithm is largely compensated by the gains obtained in the long run. This is still more effective if a first level of preconditioning is applied to cluster better the spectrum of the iteration matrix.

7 Concluding Remarks

The BlockCGSI algorithm computes a *near*-invariant subspace, associated with the smallest eigenvalues in $M^{-1}A$, which combines the subspace inverse iteration and a stabilized version of the block Conjugate Gradient algorithm. The main focus in this work was the control of the accuracy when solving the system with multiple right-hand sides at each inverse iteration, and the good agreement of the stopping criterion used in the blockCG iteration with the measure of convergence of the inverse iteration itself. Similarly to other inexact inverse iteration analysis (for instance [9, 11, 20]), we analyze the inner-outer iteration in the blockCGSI algorithm, and we propose to measure the residuals of the system through a Rigal–Gaches type of backward error. This measure enables the control of the absolute eigenvalue error of the inverse iteration, at the same time that the system is solved. The control is even more effective at the first inverse iteration if the starting vectors are previously filtered with Chebyshev polynomials. We also derive an expression, linked to the proposed residual measure, that indicates when the inner iteration must be stopped if we want to recover the same type of convergence as in the exact inverse iteration. Based on the asymptotic behavior of this expression, we suggest how to avoid unnecessary extra computational work in the blockCG inner iteration.

We also investigated some particular techniques, like the Chebyshev filtering of the random generated vectors, and a form of dynamic adjustment of the dimension of the

current subspace at each inverse iteration. The experiments indicated that Chebyshev filtering is useful to reduce the total number of inverse and blockCG iterations, and consecutively to reduce the total amount of work. The *"sliding window"* technique is helpful to make the algorithm flexible and robust. Some of the good features of the BlockCGSI algorithm (see Algorithm 1) also yield in the easy control of the memory requirements as well as the a priori control of the accuracy.

Once we have computed the spectral information associated with the smallest eigenvalues, we experiment different strategies for improving the consecutive solution of linear systems with the Conjugate Gradient algorithm. In that respect, we have focused on two closely related approaches: (1) deflating the eigencomponents associated with the smallest eigenvalues with an appropriate starting guess, or (2) using the SLRU preconditioner that shifts these eigenvalues away from zero. The latter appeared to be numerically more stable, achieving linear convergence, even when the pre-computed spectral information was obtained with low accuracy. Nevertheless, the first approach is less expansive in terms of computational cost, and is a preferable option if the multiples systems are solved with a not very small stopping criterion.

The experiments show that, if the spectrum is previously clustered, with the help for instance of a first level of preconditioning, the strategy is very efficient in the reduction of the total cost of solving consecutive linear systems with changing right-hand sides. The extra work needed to compute the spectral information is payed back after a small number of consecutive solutions.

The two-phase strategy is also very effective in other applications. In previous work [19], were it was used to accelerate the simulation of the flow around an airplane wing, we have verified that the reduction of the total amount of computational costs can reach 70%.

References

1. Parlett, B.N.: The Symmetric Eigenvalue Problem. SIAM, Philadelphia (1998)
2. O'Leary, D.P.: The block conjugate gradient algorithm and related methods. Linear Algebra Appl. **29**, 293–322 (1980)
3. Arioli, M., Duff, I., Ruiz, D., Sadkane, M.: Block Lanczos techniques for accelerating the block cimmino method. SIAM J. Sci. Stat. Comput. **16**(6), 1478–1511 (1995)
4. Balsa, C., Daydé, M., Palma, J.M.L.M., Ruiz, D.: Monitoring the block conjugate gradient convergence within the inexact inverse subspace iteration. In: Wyrzykowski, R., Dongarra, J., Meyer, N., Wasniewski, J. (eds.) Parallel Processing and Applied Mathematics. Lectures Notes in Computer Science - PPAM05, LNCS 3911, pp. 494–504. Springer, Berlin (2006)
5. Arioli, M., Ruiz, D.: Block conjugate gradient with subspace iteration for solving linear systems. In: Margenov, S., Vassilevski, P. (eds.) Iterative Methods in Linear Algebra, Second IMACS Symposium on Iterative Metohds in Linear Algebra, pp. 64–79. Blagoevgrad, Bulgaria (1995)
6. Giraud, L., Ruiz, D., Touhami, A.: A comparative study of iterative solvers exploiting spectral information for SPD systems. SIAM J. Sci. Comput. **27**(5), 1064–8275 (2006)
7. Carpentieri, B., Duff, I., Giraud, L.: A class of spectral two-level preconditioners. SIAM J. Sci. Comput. **25**(2), 749–765 (2003)
8. Lai, Y.L., Lin, K.Y., Lin, W.W.: An inexact inverse iteration for large sparse eigenvalue problems. Numer. Linear Algebra Appl. **4**(5), 425–437 (1997)

9. Smit, P., Paardekooper, M.H.C.: The effects of inexact solvers in algorithms for symmetric eigenvalue problems. Linear Algebra Appl. **287**, 337–357 (1999)
10. Golub, G.H., Ye, Q.: Inexact inverse iteration for generalized eigenvalue problems. BIT **40**, 671–684 (2000)
11. Simoncini, V., Eldén, L.: Inexact Rayleigh quotient-type methods for eigenvalue computations. BIT **42**, 159–182 (2002)
12. Berns-Mueller, J., Graham, I. G., Spence, A.: Inexact inverse iteration for symmetric matrices. Linear Algebra and its Appl. **416**(2–3), 389–413 (2006)
13. Notay, Y.: Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem. Numer. Linear Algebra Appl. **9**, 21–44 (2002)
14. Dongarra, J.J., Ducroz, J., Hammarling, S., Duff, I.: A set of level 3 basic linear algebra subprograms. ACM Trans. Math. Softw. **16**(1), 1–28 (1990)
15. Rigal, J.L., Gaches, J.: On the compatibility of a given solution with the data of a linear system. J. ACM **14**(3), 543–548 (1967)
16. Hageman, L.A., Young, D.M.: Applied Iterative Methods. Academic Press, New York (1981)
17. Van der Sluis, A., van der Vorst, H.A.: The rate of convergence of Conjugate Gradients. Numer. Math. **48**(5), 543–560 (1986)
18. Dongarra, J.J., Ducroz, J., Hammarling, S., Hanson, R.: An extended set of Fortran basic linear algebra subprograms. ACM Trans. Math. Softw. **14**, 1–17 (1988)
19. Balsa, C., Braza, M., Daydé, M., Palma, J.M.L.M., Ruiz, D.: Improving the numerical simulation of an airflow problem with the BlockCGSI algorithm. In: Daydé, M., Palma, José M.L.M., Coutinho, Álvaro L.G.A., Pacitti, Esther, Correia Lopes, J. (eds.) High Performance Computing for Computational Science - VECPAR 2006. LNCS 4395, pp. 281–291. Springer, Berlin (2007)
20. Notay, Y.: Convergence analysis of inexact Rayleigh quotient iteration. SIAM J. Matrix Anal. Appl. **24**(3), 627–644 (2003)

Location Around Big Cities as Central Places

Fernando Barreiro-Pereira

Abstract Space behaviour is very close to imperfect competition: agglomeration economies lead to increasing returns to scale and physical distance between economic agents and markets causes horizontal product differentiation and prices discrimination in goods markets and in land as a good and as an input. An important target in macroeconomics is the money market analysis, but money market is generally not considered in microeconomic models and hence neither in most spatial models although it affects the location. The main aim of this paper is to introduce the money market in a general equilibrium model to explain the location of consumers and producers around a big monocentric city, where consumers choose optimal quantities of consumption goods, money, land and transport, and households and firms can rationally choose their location in relation to the central market. Results for firms indicate that their locations are generally situated beyond those of households with respect to the central business district, depending on their land needs.

Keywords Central places theory · New economic geography · Spatial general equilibrium · Location theory · Demand for money · Socially efficient transportation rates

1 Introduction

At the end of the XX Century, a great part of the global economic activity is mainly developed around metropolitan areas, which are converting in big cities that enter in some specialization and competition processes among them for the economic power. The 600 main cities produce 80 percent of global GDP in 2016. The last decades have witnessed the emergence and the never seen growth of a number of big cities: while in 1950 there were 2 mega-cities with more than 9 million inhabitants, in 2016 there were 49, many of them being located in less developed countries. However it happens that the size seems to be neither a necessary nor a sufficient condition for obtaining the status of global city. A condition to be a global city is the access to the economic

F. Barreiro-Pereira (✉)
Department of Economic Analysis, National University for Distance Education (UNED Univ.), Paseo Senda Del Rey 11, 28040 Madrid, Spain
e-mail: fbarreiro@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_4

power, following the global city hypotheses re-elaborated by Sassen [35]. In most cases the structure of these large cities is monocentric, although in other cases they are composed of several centers forming a megalopolis. The causes of the increase in the size of the cities are related to the levels of per capital income and employment, accompanied by low transport costs. The approach proposed in this work to explain the location of households and firms around big metropolitan areas is a spatial general equilibrium model which embodies consumption goods, money, inputs, transport and land as in Isard [24], but in an imperfect competitive framework as in Thisse [39] or Fujita et al. [16]. In the present research we assume the existence of a central place with a central market in their central business district (CBD) immersed in a metropolitan area without migration flows; the background of this research is related to the works of von Thünen [41], Alonso [1], Muth [29] in relation to the location theory, bid-rent and urban residential land use around a monocentric city; Herbert and Stevens [21] for housing market analysis and Christaller [7], Lösch [27] for their central place theory. The preliminary works on central place theory focused on the identification of the geometric conditions that make possible an overlap of regular structures, but with too few microeconomic foundations to explain the grouping of households and firms. Early contributions that analyze spatial distribution from powerful microeconomic foundations use partial equilibrium models. In this sense, one of the first economic contributions to the theory of central places is due to Eaton and Lipsey [8] who, by assuming concurrence of multipurpose shopping which results in demand externalities, develop a model of spatial competition which gives rise to the emergence of central places. They consider that firms occupy no space along one-dimensional market. Quinzii and Thisse [33] consider that consumers are distributed along a circle, but they use a similar approach to determine the location of firms that maximizes well-being, demonstrating that without some government intervention market equilibrium may be inefficient. Fujita [12] also considers a long-narrow country represented by one-dimensional unbounded location space, where the firms are in monopolistic competition. The resulting model resembles more that of Pred [31] as that of Chamberlin [5]. Spatial agglomeration is generated in this model through the product variety in consumption goods. Considering an economy in a boundless one-dimensional space where the firms have increasing returns, Mori [28] shows that declines in transport costs of consumption goods promote the dispersion of industrial activities, resulting in the formation of a megalopolis. In the same sense and under a long narrow spatial economy, Fujita and Mori [14] extend the monocentric spatial-economy model of Fujita and Krugman [11] to a multi-city model whenever the population exceeds a critical value. In the long-run the spatial system of the economy will approach a highly regular central place system. Later, Hsu [22] analyzes the city size distribution along a linear segment by means of a Bertrand oligopolistic competition model in partial equilibrium in which cities of different sizes serve different functions in the economy, specifying the conditions under which the central place hierarchy becomes a fractal structure. Hsu and Holmes [23] used an approach similar to Quinzii and Thisse [33], but extending it to several sectors. In relation to the spatial general equilibrium models, Fujita et al. [15] use the framework of the new economic geography to generate a hierarchical city system

by developing a general equilibrium model for a multi-industrial economy located in an unlimited one-dimensional space, as an extension of the Krugman [26], Fujita and Krugman [13] core-periphery model; in this model, as the population increases a more or less regular hierarchical central place system emerges. Tabuchi and Thisse [38] also use a model of general equilibrium with monopolistic competitive markets for industrial sectors with firms that present increasing returns to scale. In the model the spatial economy described by a circumference of length one and the number, size and location of cities are determined endogenously as a result that the principle of hierarchy of the central places theory is triggered by the fall of the transport costs. Chen and Partridge [6] from Glaeser [17], apply a general equilibrium model with firms that present constant returns to scale in a central place framework of spread-backwash adapted to China by incorporating elements from the new economic geography. Results indicate that market potential in China's mega-cities is inversely related to growth for smaller cities and rural communities, that is, growth in the mega-cities may reduce growth elsewhere.

The objective of the present research is not to verify that the principle of hierarchy is fulfilled as it happens in most of the works on central places above mentioned including those of Eaton and Lipsey [8], Stahl [37], although this could be deduced from the model and its assumptions. The real purpose of this paper is to explain the location of consumers and producers with respect to the center of a central place, where we assume that the central market is located. Consumers maximize their utility conditioned by two constraints: a budget and a temporary one, while producers maximize profit. The results indicate that consumers are located at a distance from the central market that depends on whether the land size of their home and their working time are stipulated or not. If both situations occur, the consumers will be placed inside a circular crown with center in the central market, whose radii are the two distances to the center. Unlike Eaton and Lipsey [8], in the present research firms have increasing returns to scale and occupy space. The location of producers depends on the amount of land they need to produce consumer goods and they may also be located into the circular ring area grouping together to maintain increasing returns to scale. This can cause negative externalities that are likely to provoke consumers to cluster at nodes opposed to firms, forming two inverted regular triangular networks around central place when transport rates are efficient. Space is incorporated into this model from three points of view: First, land is considered as a good for the consumers and as an input to the firms and we suppose that the quantity of land around the central market is unlimited. Second, consumers have to travel from their homes to the CBD in order to purchase goods and producers must transport their goods from their factories or warehouses to the CBD in order to sell them to consumers. Therefore our model contemplates two kinds of transport: passenger and freight, and there are respectively two kinds of unitary prices in the transport system. Third, a particular novelty of this research is the consideration of money in the consumer utility function, which implies that the transaction money demand may have influence on the location decisions. We assume in the model that money, transport and production factors markets are in perfect competition, while the consumer goods and consumption of land markets are in imperfect competition. The work is structured as follows: Sect. 2 describes

some hypotheses about the economy. In Sect. 3 we present the consumers choice and the effects of the money market on location. Section 4 describes the case of socially efficient tariffs for passenger transport. In Sect. 5 is presented the choice of firms and their location. Section 6 describes the distribution of the market areas and, finally, Sect. 7 draws some relevant conclusions on this work.

2 The Economy

In our economy, we assume that F firms produce n consumption goods $(X_1, X_2, \ldots, X_n)$ using m inputs $(L_1, L_2, \ldots, L_m)$, where labour and physical capital are included. The vector of goods prices is $\mathbf{P} = (P_1, P_2, \ldots, P_n)$, and the vector of inputs prices is $\mathbf{W} = (W_1, W_2, \ldots, W_m)$. Consumption goods markets are in imperfect competition. The quantity of input (j) purchased by a firm (f) to all households is L_j^f; and X_i^f is the quantity of product (i) sold by the firm (f).The quantities of goods sold by a firm (f) is $\mathbf{X}^f = (X_1^f, X_2^f, \ldots, X_n^f)$ and the quantities of inputs purchased by a firm (f) is $\mathbf{L}^f = (L_1^f, L_2^f, \ldots, L_m^f)$. Moreover, in this economy there are H households possessing all inputs and sell them to the firms. Each household provides in his/her working time several quantities of all production factors $\mathbf{L}^h = (L_1^h, L_2^h, \ldots, L_m^h)$ to several firms and the income coming from selling it is applied by the households to buy some goods, being X_i^h the quantity of good (i) purchased by a household (h). In addition, we suppose that the households have some property rights on the firms (s^{hf}), like profit sharing, for instance. In this economy we assume, to simplify, the existence of one public firm which is the only land proprietary; land (Z) is rented monthly to the households and firms. Land market is in imperfect competition, either as a good for consumers or as an input for firms, because land is horizontally differentiated by means of the distance to central market. The unit price of land (q) is not constant because its value depends on the distance to CBD. Following Muth [29] there is a distance decay in land prices from the central market to the periphery because land is scarcer in the center than in the periphery. Land causes non-convexities in preferences and to avoid the difficulties that this causes, we will assume: (1) Agents are unable to consume land at more than one location at a time; (2) The economy has a large number of households. Apart from land, the other non-produced good in this economy is money. The incorporation of money in a general equilibrium model implies some problems because real money demand function is homogeneous of degree one, whereas the demand functions of the consumption goods, consumption of land and leisure are homogeneous of degree zero in prices and income. This is related to the fact that in a general equilibrium model goods that transfer wealth over time, such as money, cannot be mixed with goods that do no transfer it. That is, we cannot mix goods as flows with goods as stock in a general equilibrium model. This dichotomy is solved by Kuenne [25] by assuming two kinds of money: stock and flow. Money as a flow (M) does not embody any utility; its market is perfect and its price is unique and it makes the role of numerary in our general equilibrium

model. We assume that the stock money market is exogenous to the model and its price is the common nominal interest rate. Saving (S) is also considered as a good in the consumer problem, whose price (r) is the real interest rate. This economy has a transport system. Firms transport their commodities contracting some transportation companies, which only make journeys from the firms to the central market. By simplicity, the transport of the raw materials from the origin to the factories is not considered in this problem. We also suppose that the origins of raw materials are outside the metropolitan area. Travel to work are not considered unless they are to work in the central market. We assume that transport expenditure is linear according to the distance covered although this may cause congestion. There are two kinds of tariffs: One for freight transport (t_m) and the other one for passengers transport (t_p) paid by the consumers who make trips to the center and return to home; we assume that these tariffs are competitive. Apart from that, we assume that the representative consumer has symmetrical preferences over all differentiated products, and hence the H households have the same preferences and identical utility functions. This fact causes identical demand functions. By assuming representative agents, the total consumption made by the consumers will be nX_iH quantities of goods. If an f firm is representative, each firm will transport once a day to central market $nX_i(H/F)$ quantities of commodities produced.

Every consumer will buy nX_i quantities of goods that will be transported to home from the central market once a day; following Small [36] the cost of this transport is $t_p\delta^h$ plus the cost of the journey from home to central market, that is, $2t_p\delta^h$, being δ^h the home-market distance for each consumer. The transport cost of each group of nX_i amounts of goods is $t_m\delta^f$ assuming that only travels from the firms to the central market are paid but not their empty transport returns; δ^f is the firm-market distance and t_m the rate for freight. Finally, we can assume as known the profits sharing for consumers, s^{hf}, and the initial quantity of flow money that consumers own (M_s^h), where the subscript "s" stands for supply. We also suppose that the sum of the initial quantities of flow money should be identical to the final sum of flow money demanded (M_d). Consumers can invest some previous savings (S^h) in the actual period, but only in flow money form, and therefore, $M_s^h = M^{h*} + rS^h$, where r is the saving price or real interest rate and M^{h*} the initial endowment of liquid money without consider money coming from savings.

3 Consumers Choice and Location

Under the hypothesis above mentioned and taking a working day as a unit of time, the budget constraint for a consumer h has the following form:

$$Y^h \equiv \sum_{j=1}^{m}\left(W_jL_j^h\right) + \sum_{f=1}^{F}\left(s^{hf}B^f\right) + M_h^* + rS_h = \sum_{i=1}^{n}\left(P_iX_i^h\right) + q^h(\delta^h)Z^h + M_d^h + 2t_p\delta^h \tag{1}$$

or in matrix notation, the total income received by the consumer is $Y^h = \mathbf{WL}^h + \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h}$ and this total income is spent by the consumer in the purchase of all goods, including transportation:

$$Y^h \equiv \mathbf{WL}^h + \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h} = \mathbf{PX}^h + q^h(\delta^h)Z^h + M_d^h + 2t_p\delta^h \tag{2}$$

M_d^h is the final quantity of flow money demanded, which contains the actual savings to be invested it in future periods. $\mathbf{W}$ contains the input prices in monetary terms, and $\mathbf{B}^f$ are the firm's profits; Y^h is the nominal per capital income. In the Eq. 2 the underlined terms $\underline{\mathbf{s}^{hf}\mathbf{B}^f}$ and $\underline{M_s^h}$ mean that they are known in advance. In the consumer choice problem, the axiom of rationality implies the maximization of a particular utility function. The need to avoid Say's law, or the result that the supply of goods produced by firms automatically generates its own demand, imply that the agents can choose among several goods: in the standard macroeconomics model the choice is between consumption and saving. In other models, the choice occurs between produced goods and a non-produced good, as in Hart [19]. Here, the consumer chooses among consumption goods, land, money, and leisure. For a specific consumer, the utility function may be: $U^h = U^h(X_1^h, X_2^h, \ldots, X_n^h, Z^h, M_d^h, \Omega^h)$. Following Fujita [10] in his time extended model, the consumer is also subjected to a temporary constraint which explains the total time used by the consumer ($\underline{T}$ during a day, for example) as the sum of leisure time (Ω^h), plus the time of transport, plus the time to work (labour) and for supplying directly or indirectly other production factors ($\mathbf{L}^h$) to the firms. Being δ^h the distance from the consumer to the central market, each consumer travel this distance twice a day to purchase the $\mathbf{X}^h$ goods; if the average speed of passengers transportation is V, then: $V = 2\delta^h/\tau$, where τ is the time for transport; hence, we can express leisure as: $\Omega^h = \underline{T} - \mathbf{L}^h - 2\delta^h/V$, being the time $\underline{T}$ known and fixed.

On the other hand, money is directly incorporated into the utility function and in a cash in advance constraint, according to the technology of transactions. In order to warrant the correct aggregation of goods and agents, following Greenhut et al. [18], the utility function must be quasi-concave, and homogeneous of degree one in consumption, land, leisure and real money balances, as well as multiplicatively separable in a per capital composite consumption good (C), consumption of land (Z), leisure (Ω) and real money balances ($M_d = DM/P$), being DM the monetary demand of money as flow and P an index that reflects the level of prices represented by the average prices:

$$P = \left(\frac{1}{1+n+m+H+F}\right)\left(q_0 + \sum_1^H q^h + \sum_1^F q^f + \sum_1^n P_{X_i} + \sum_1^m W_j\right) \tag{3}$$

where q_0 is the land price in the city center; q^h and q^f are the land prices paid respectively by consumers and producers; P_{X_i} are the prices of consumption goods and W_j the prices of production factors.

We consider C as a composite good of all consumption goods X_i, and we assume that each consumer h demands $n \times i$ quantities of composite good C^h. We compute the C^h index, as follows:

$$C^h = n^{\frac{1}{1-\theta}} \left(\sum_{i=1}^{n} \left(X_i^h \right)^{\frac{\theta-1}{\theta}} \right)^{\frac{\theta}{\theta-1}} \tag{4}$$

where all goods have been placed in symmetrical form, being θ a parameter which reflects the constant elasticity of substitution between consumption goods. If θ is large then consumption good are closed substitutes. The value of θ must be $\theta > 1$, because it is necessary to warranty that the price-elasticity of individual demand functions cannot be less than one, to obtain an equilibrium. We take the following index as the price of the composite consumption good (P_C):

$$P_C = \left(\frac{1}{n} \sum_{i=1}^{n} P_{X_i}^{1-\theta} \right)^{\frac{1}{1-\theta}} \tag{5}$$

3.1 Optimal Distance for Consumers: General Case

We assume that consumers in the economy are rational and efficient, that is, they carry out a conditional maximization of utility. The utility function for each consumer could then be as follows:

$$U^h = Z^{\alpha} C^{\rho} \left(M_d^h \right)^{\gamma} \Omega^{\beta} \tag{6}$$

where the exponents of the arguments must be $\alpha + \rho + \gamma + \beta = 1$. The two constraints of the consumer's problem are:

(1) Budget constraint:

$Y^h = \mathbf{WL}^h + \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h} = P_c C^h + q^h(\delta^h) Z^h + M_d^h + a\delta^h$, where $a\delta^h$ is the transport cost ($a = 2t_p$). $\overline{\text{All}}$ the terms of this budget constraint are measured in monetary terms.

(2) Time constraint:

$\underline{T} = \mathbf{L}^h + \Omega^h + b\delta^h$, that is, the consumer distributes his/her unit of time $\underline{T}$ (one working day for example) as time dedicated to the work ($\mathbf{L}^h$), more time dedicated to the leisure (Ω^h), more time dedicated to the transport ($b\delta^h$),where $b\delta^h$ is the transport time ($b = 2/V$) and V is the average speed of passenger transport in the city. All the terms of this time constraint are measured in time, including the vector of production factors ($\mathbf{L}^h$) employed by a household h. Being then $\mathbf{L}^h = \underline{T} - \Omega^h - b\delta^h$ and by replacing $\mathbf{L}^h$ into $Y^h = \mathbf{WL}^h + \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h}$, so-calling Y_n the non- labour income ($Y_n = \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h}$), we will have the following composed constraint, whose terms are in monetary $\overline{\text{units}}$:

$$Y_n + W\underline{T} - (Wb + a)\,\delta^h = P_C C^h + q^h\left(\delta^h\right) Z^h + M_d^h + W\Omega^h \tag{7}$$

where the left side of the equation 7: $Y_n + W\underline{T} - (Wb + a)\,\delta^h$ is the available, or disposable income net of transport costs, with which the consumer can buy all the goods located on the right side of the equation, that is, all goods except transport. The left side of the Eq. 7 is called in the paper as $I\left(\delta^h\right)_1$, that is, $I\left(\delta^h\right)_1 = Y_n + W\underline{T} - (Wb + a)\,\delta^h$, which as we can see it is in principle dependent on the distance to the central market (δ^h). In the Sects. 3.3 and 3.4, depending of the cases, this disposable or net income has other different composition that here in Sect. 3.1, being then called respectively $I\left(\delta^h\right)_2$ and $I\left(\delta^h\right)_3$. By solving the maximization of the utility submitted to the constraint (7), we have:

$$C^h = \frac{\rho}{P_C} I\left(\delta^h\right)_1;\; M_d^h = \gamma\, I\left(\delta^h\right)_1;\; Z^h = \frac{\alpha}{q\left(\delta^h\right)} I\left(\delta^h\right)_1;\; \Omega^h = \frac{\beta}{W} I\left(\delta^h\right)_1 \tag{8}$$

These results are the Marshallian demand functions of the consumption goods, land and leisure and all are homogeneous functions with degree zero, except for money (degree one). By substituting the Marshallian demands into the direct utility function, we obtain the indirect utility function:

$$v = \Phi\left(\frac{I\left(\delta^h\right)_1}{\left[q\left(\delta^h\right)\right]^{\alpha}}\right) \tag{9}$$

where $\Phi = \left[\alpha^{\alpha}\left(\frac{\rho}{P_C}\right)^{\rho}\gamma^{\gamma}\left(\frac{\beta}{W}\right)^{\beta}\right] \neq 0$. By equaling to zero the derivative of v with respect to δ^h we can find the optimum value for δ^h whenever the indirect utility function v is concave with respect to δ^h ($v''_{\delta\delta} < 0$):

$v'_{\delta^h} = \frac{\Phi}{\left[q\left(\delta^h\right)\right]^{2\alpha}}\left(\left[q\left(\delta^h\right)\right]^{\alpha}\frac{dI\left(\delta^h\right)_1}{d\delta^h} - \alpha\left[q\left(\delta^h\right)\right]^{\alpha-1}\frac{dq\left(\delta^h\right)}{d\delta^h} I\left(\delta^h\right)_1\right) = 0$, then,

$\left[q\left(\delta^h\right)\right]^{\alpha}\frac{dI\left(\delta^h\right)_1}{d\delta^h} = \alpha\left[q\left(\delta^h\right)\right]^{\alpha-1}\frac{dq\left(\delta^h\right)}{d\delta^h} I\left(\delta^h\right)_1$, that is, $\frac{\frac{dI\left(\delta^h\right)_1}{d\delta^h}}{I\left(\delta^h\right)_1} = \alpha\frac{\frac{dq\left(\delta^h\right)}{d\delta^h}}{q\left(\delta^h\right)}$, but $I\left(\delta^h\right)_1 = Y_n + W\underline{T} - (Wb + a)\,\delta^h$ and $\frac{dI\left(\delta^h\right)_1}{d\delta^h} = -(Wb + a)$. Replacing it, we have:

$\frac{-(Wb+a)}{I\left(\delta^h\right)_1} = \alpha\frac{\frac{dq\left(\delta^h\right)}{d\delta^h}}{q\left(\delta^h\right)}$; and hence:

$$\frac{dq}{q} = \frac{-(Wb + a)}{\alpha I\left(\delta^h\right)_1} d\delta^h \Rightarrow \alpha\int\frac{dq\left(\delta^h\right)}{q\left(\delta^h\right)} = \int\frac{-(Wb + a)}{Y_n + W\underline{T} - (Wb + a)\,\delta^h} d\delta^h \tag{10}$$

Resolving integrals (10): $\alpha\log_e\left[q\left(\delta^h\right)\right] = \log_e[Y_n + W\underline{T} - (Wb + a)\,\delta^h] + \log_e C$

where $\log_e C$ is a constant to be determined next: Being q_0 the unit land price at central market, when $\delta^h = 0$ then $q\left(\delta^h\right) = q_0$. Therefore: $\log_e C = \alpha \log_e(q_0) - \log_e[Y_n + W\underline{T}]$. Substituting $\log_e C$ by its value:
$\alpha[\log_e\left[q\left(\delta^h\right)\right] - \log_e(q_0)] = \log_e[Y_n + W\underline{T} - (Wb + a)\,\delta^h] - \log_e[Y_n + W\underline{T}]$.

And isolating δ^h from the last expression we can know the optimum radial distance from home to central market (δ_1^h):

$$\delta_1^h = \frac{Y_n + W\underline{T}}{Wb + a}\left[1 - \left(\frac{q\left(\delta^h\right)}{q_0}\right)^{\alpha}\right] = \frac{I\left(\delta^h\right)_1}{Wb + a}\left[\left(\frac{q_0}{q\left(\delta^h\right)}\right)^{\alpha} - 1\right] \tag{11}$$

where $q\left(\delta^h\right)$ is the unit land price at the distance δ_1^h in this case. The second part of Eq. (11) comes from considering that $Y_n + W\underline{T} - (Wb + a)\,\delta^h = I\left(\delta^h\right)_1$ and replacing it in the first part of the Eq. (11):

$(Wb + a)\,\delta^h = I\left(\delta^h\right)_1\left[1 - \left(\frac{q\left(\delta^h\right)}{q_0}\right)^{\alpha}\right] + (Wb + a)\,\delta^h\left[1 - \left(\frac{q\left(\delta^h\right)}{q_0}\right)^{\alpha}\right]$, isolating δ^h, we have the second expression of the optimum distance consumer's home-central market (δ_1^h), now in function of the net income $I\left(\delta^h\right)_1$.

By assuming that $(q_0/q)^{\alpha}$ can tend to one, we can develop the term $[(q_0/q)^{\alpha} - 1]$ by means of a McLaurin series (one term plus the residual) and the second part of the formulation 11 could be then expressed as follows:

$$\delta_1^h = \frac{\alpha I\left(\delta^h\right)_1}{Wb + a}\log_e\left(\frac{q_0}{q\left(\delta^h\right)}\right) \tag{12}$$

which is an implicit function that can be solved considering the existence of the money market.

3.2 Money Market Effects on Location and Optimum Land Size

From the Marshallian demand function of land as a good, we can see that the amount of land demanded is dependent of the distance to the central market. The best model to select the optimum amount of land is the Herbert and Stevens [21] model, which in a competitive land market, maximizes the total surplus subject to land and population constraints. However, our analysis considers that the land market is imperfectly competitive and we also consider that the total amount of land around the central market is unlimited; the city is expanded still land prices are zero (we assume that there is not any agriculture land around central place). These facts invalidate the strict application of the Herbert-Stevens model in our framework and its land and population constraints to select the optimum land size.

To select the optimum size of land we suppose that the consumer, at least in the short run, demands money to make his/her transactions according to the Baumol [2],

Tobin [40] model. By assuming that there are two ways of storing wealth: money and bonds and that the disposable income to buy consumption goods and land depends of the distance to the central market, following the Baumol-Tobin model for transactions money demand, the optimum amount of cash money demanded by a rational consumer who minimizes the total cost of keeping money against bonds in a given time interval, follows the square root rule:

$$M_d^h = \gamma I\left(\delta^h\right)_1 = \sqrt{\frac{cI\left(\delta^h\right)_1}{2i}} \tag{13}$$

where c is a fixed cost per transaction which is independent of the amount of bonds exchanged by money as flow. The term $I\left(\delta^h\right)_1$ is the disposable income, and i is the nominal interest rate; i is related with the real interest rate (r) by means of the Fisher [9] equation: $i = r + \pi^e$, where π^e is the expected inflation rate. Therefore, the transaction money demand M_d^h increases when the fixed cost c and nominal net income $I\left(\delta^h\right)_1$ increase, but it decreases when the interest rate i or the expected inflation rate π^e increase. From the above equation we can deduce the optimum disposable net income:

$$I\left(\delta^h\right)_1 = \frac{c}{2i\gamma^2} \tag{14}$$

which now results independent of the distance. Like $I\left(\delta^h\right)_1 = Y_n + WT - (Wb + a)\,\delta^h$, replacing $I\left(\delta^h\right)_1$ by its value in Eq. (14) and isolating the distance, we have:

$$\delta^h = \frac{Y_n + W\underline{T}}{Wb + a}\left[1 - \frac{c}{2i\gamma^2\left(Y_n + W\underline{T}\right)}\right] \tag{15}$$

When this distance is optimal, Eqs. (12) and (15) must be equal and this means that:

$$q(\delta^h) = q_0\left[\frac{c}{2i\gamma^2\left(Y_n + W\underline{T}\right)}\right]^{\frac{1}{\alpha}} \tag{16}$$

By substituting both $I\left(\delta^h\right)_1$ and $q(\delta^h)$ in the Marshallian demand function of Z^h (Eq. 8), we will obtain the optimum land size at distance δ_1^h:

$$Z_1^h = \frac{\alpha}{q_0}\left(\frac{c}{2i\gamma^2}\right)^{1-\frac{1}{\alpha}}\left(Y_n + W\underline{T}\right)^{\frac{1}{\alpha}} \tag{17}$$

We can see that the optimal amount of land purchased by the consumer is increasing with their incomes and wages, but it is inversely proportional to the unit land price in the center and the nominal interest rate. The latter can be clearly seen in the

case of mortgages. From the land demand function (Eq. 8), we can also see that the total expenses in land at the optimum distance result independent of the distance, but it depends on the interest rate:

$$Z^h q(\delta^h) = \alpha I \left(\delta^h\right)_1 = \frac{\alpha c}{2i\gamma^2} = @ \tag{18}$$

where @ is a constant if the nominal interest rate (i) is constant. Replacing (18) in the expression (12), we obtain the optimal distance between the consumer and the central market:

$$\delta_1^h = \frac{@}{2(\frac{W}{V} + t_p)} \ln\left(\frac{q_0 Z^h}{@}\right) = \frac{@}{2(\frac{W}{V} + t_p)} \ln\left(\frac{q_0}{q_1}\right) \tag{19}$$

because by Eq. (18) we know that $@ = Zq$; q_1 is the unit land price at δ_1^h. We can see that the optimal distance to which a rational consumer is located with respect to the center tends to be greater the greater unit land price in the center, the amount of land acquired and the speed of passenger transport; and the distance tends to be smaller the higher the consumers wage, the passenger transport rates and the higher the interest rate and the expected rate of inflation. The empirical verification of the impact of the money market on space is outside this research. By Eqs. (12) and (19) we can deduce the expression of the law of land prices from the central market to the periphery, for a generic unit land price q^h to be paid by a consumer h located at distance δ^h from the CBD:

$$q^h = q_0 e^{\frac{-2(\frac{W}{V}+t_p)}{@}\delta^h} = q_0 e^{\frac{-2(Wb+a)}{\alpha I\left(\delta^h\right)_1}\delta^h} \tag{20}$$

On the other hand, by substituting the value of the leisure time selected at the optimum distance (Eq. 8) into the time constraint we can also deduce the optimum labour time for each consumer, once considered the Eq. (18):

$$L_1^h = \underline{T} - \frac{c}{2i\gamma^2}\left[\frac{\alpha}{W + Vt_p} \log_e\left(\frac{2i\gamma^2 q_0 Z_1^h}{\alpha c}\right) + \frac{\beta}{W}\right] \tag{21}$$

Here we can see that for a consumer the optimal amount of time devoted to work tends to be greater the higher his/her wage and the interest rate.

3.3 Consumers Choice Under Fixed Labour Time and Land Size

Normally the time of labour tends to be fixed depending of each job. For example, a certain number of consumers work 35 hours a week. We can assume that, at least for a part of the consumers, labour time keep fixed and its value is $L^h = \underline{L}^h$. Besides

that, as a result of urban housing construction plans, a large part of consumers end up accepting a medium-sized home, so for these consumers we will assume that their land size is fixed ($Z^h = \underline{Z}^h$) whatever their distance from the CBD. This situation is different from the previous general case where consumers-workers could freely choose the optimal amounts of work and land.

With respect to the utility maximization in the general case (Eq. 6), under these two assumptions the problem of the consumer is now: to maximize an utility function whose arguments are only consumption goods and real money balances, because the other goods are fixed, subject to a constraint where land and labour are fixed:

$$\max\langle U = (C^h)^{\rho}(M_d^h)^{\gamma}\rangle \tag{22}$$

subject to: $Y_n + W\underline{L}^h = P_C C^h + q^h\left(\delta^h\right)\underline{Z}^h + M_d^h + \alpha\delta^h$ or rearranging the composed consumer's constraint:

$$I\left(\delta^h\right)_2 = Y_n + W\underline{L}^h - \left[\alpha\delta^h + q^h\left(\delta^h\right)\underline{Z}^h\right] = P_C C^h + M_d^h \tag{23}$$

where now $[I\left(\delta^h\right)_2]$ is the disposable income to buy the C and M goods. By assuming that $\rho + \gamma = 1$, the indirect utility function is:

$$v = \gamma^{\gamma}\left(\frac{\rho}{P_c}\right)^{\rho} I\left(\delta^h\right)_2 = \gamma^{\gamma}\left(\frac{\rho}{P_c}\right)^{\rho}\left(Y_n + W\underline{L}^h - \left[\alpha\delta^h + q^h\left(\delta^h\right)\underline{Z}^h\right]\right) \tag{24}$$

The apparent result of maximizing v respect δ^h is:

$$\frac{\partial\left[q^h\left(\delta^h\right)\right]}{\partial\delta^h} = \frac{-a}{Z^h\left(\delta^h\right)} \Longrightarrow q^h\left(\delta^h\right) = q_0 - \left(\frac{a}{\underline{Z}^h}\right)\delta^h \tag{25}$$

This is the land prices law in this particular case, which differs from that of the general case. In any case, the radial land prices law must be unique around the central place, since for each location there is only a single unit price of land and the highest of them is q_0, while the rest of the prices form circular level curves around the center. One can discuss which of these two land prices law should prevail or whether the final land prices law should be an average of the two laws weighted by the number of corresponding consumers affected. In the absence of concrete data we have chosen in this work to suppose that the prevailing land prices law is the one that was previously established by the consumer behaviour in the general case (Eq. 20), where there are no restrictions in the amount of land neither in the working time, because we assume that most consumers can choose affordable amounts of all goods.

Therefore, maximizing the indirect utility function (Eq. 24) with respect to the distance, i.e., deriving it and equating it to zero, we have:

$$(\underline{Z}^h)q_0\left(\frac{Wb + a}{@}\right)e^{-\left(\frac{Wb+a}{@}\right)\delta_2^h} = a \tag{26}$$

where δ_2^h is the optimal distance from home to the central market for the consumers with restrictions in land and labour time. By taking natural logarithms in (26), we will have a new relocation for these consumers:

$$\delta_2^h = \frac{@}{Wb + a} \log_e \left(\frac{(Wb + a)\, \underline{Z}^h q_0}{\alpha @} \right) \tag{27}$$

At this distance, the consumer also minimizes the transportation and land expenses for a given land size. To prove this, we will operate following the minimum expenditure approach, knowing that in our problem $a = 2t_p$ and $b = 2/V$ and assuming that the land size obtained ($\underline{Z}^h$) by the consumer is a determined fixed amount that satisfies the consumer's tastes. The minimum expenses condition will be then:

$$\min_{\delta^h} \langle \underline{Z}^h q_0 e^{-2[(W/V)+t_p](\delta^h/@)} + 2t_p \delta^h \rangle \tag{28}$$

conditioned to the optimum land prices trajectory (Eq. 20). By deriving (28) with respect to distance and equating it to zero, we have the following expression[1]:

$$-2\left(\underline{Z}^h q_0/@\right)\left[(W/V) + t_p\right] e^{-2[(W/V)+t_p](\delta^h/@)} + 2t_p = 0 \tag{29}$$

and taking natural logarithms, we obtain the new optimum distance from these consumers to the central market:

$$\delta_2^h = \frac{@}{2(\frac{W}{V} + t_p)} \log_e \left[\frac{\underline{Z}^h q_0}{@} \left(1 + \frac{W}{V t_p} \right) \right] = \frac{@}{2(\frac{W}{V} + t_p)} \log_e \left[\frac{q_0}{q_2} \left(1 + \frac{W}{V t_p} \right) \right] \tag{30}$$

Expression identical to (27), where q_2 is the unit land price at δ_2^h. It is very important to observe now that if the consumer minimizes the land and transport costs, keeping a fixed land size, he/she will be re-located farther from central market than in the previous situation (general case), as we can see if we compare expressions (19) and (30).

3.4 Transport Time as a Part of Leisure or Labour

On the other hand ,some consumers include the time spent in transport to the CBD within the leisure time or also working time. That is to say, with respect to the general case, the time devoted to transport is removed from the time constraint of the consumer because that time is absorbed in leisure or also in the work time. The rational choice of these consumers will then be:

[1]Expression called "Muth Condition", from Muth [29]: At the equilibrium location the marginal transport cost equals the marginal land cost saving.

$$\max\langle U^h = (Z^h)^\alpha (C^h)^\rho (M_d^h)^\gamma (\Omega^h)^\beta \rangle \tag{31}$$

subject to $\underline{T} = \mathbf{L}^h + \Omega^h$ and to:

$Y^h = \mathbf{WL}^h + \underline{\mathbf{s}^{hf}\mathbf{B}^f} + \underline{M_s^h} = P_c C^h + q^h(\delta^h) Z^h + M_d^h + 2t_p\delta^h$ By introducing the first in the second consumer's constraint:

$$Y_n + WT - a\delta^h \equiv I\left(\delta^h\right)_3 = P_C C^h + q^h\left(\delta^h\right) Z^h + M_d^h + W\Omega^h \tag{32}$$

The indirect utility function in this problem is:

$$v = \left[\alpha^\alpha \left(\frac{\rho}{P_C}\right)^\rho \gamma^\gamma \left(\frac{\beta}{W}\right)^\beta\right] \left(\frac{I\left(\delta^h\right)_3}{\left[q\left(\delta^h\right)\right]^\alpha}\right) \tag{33}$$

being $I\left(\delta^h\right)_3 = Y_n + WT - a\delta^h$. By maximizing v respect to the distance δ^h we obtain a new optimal distance to central market for these consumers:

$$\delta_3^h = \frac{I\left(\delta^h\right)_3}{a}\left[\left(\frac{q_0}{q\left(\delta^h\right)}\right)^\alpha - 1\right] \simeq \frac{\alpha I\left(\delta^h\right)_3}{a} \log_e\left(\frac{q_0 Z^h}{q\left(\delta^h\right) Z^h}\right) \tag{34}$$

where $I\left(\delta^h\right)_3$is the disposable income in this case, which is different than $I\left(\delta^h\right)_1$: $I\left(\delta^h\right)_1 \neq I\left(\delta^h\right)_2 \neq I\left(\delta^h\right)_3$. Following the Eqs. (7), (18) and (32): $\alpha I\left(\delta^h\right)_1 \equiv @ = \alpha I\left(\delta^h\right)_3 + \alpha W b \delta_1^h$ and hence $\alpha I\left(\delta^h\right)_3 \neq @$. The demand function of land in the present case is $Z^h q(\delta^h) = \alpha I\left(\delta^h\right)_3$ and calling $\alpha I\left(\delta^h\right)_3$as $® \neq @$, the Eq. (34) can be written as follows:

$$\delta_3^h = \frac{®}{2t_p} \ln\left(\frac{q_0 Z^h}{®}\right) \tag{35}$$

Rearranging the Eq. (35), we have the following expression of the land prices law in this case:

$$q^h = q_0 e^{\frac{-2t_p}{®}\delta^h} \tag{36}$$

which supplies one different land prices law with respect to the general law contained in the Eq. (20), depending on the relationship between expenditures on land ® and @.

3.5 *Transport as Leisure or Work and General Land Prices Law*

We now assume that the consumers of the previous case are forced to accept the general land prices law expressed in Eq. (20), instead of governing alone by the specific law shown in Eq. (36). This is, as above mentioned, because when there are coexistence of several possible radial laws of land prices, we assume the prevalence

of the law that derives from the rational behaviour of the consumer when he/she has no restrictions in land or in working time. So that around the central market there is only one radial land prices law. The land price-radial distance gradient in our general case can be calculated from the expressions (18) and (10) as follows:

$$\frac{dq}{q} = \frac{-(Wb+a)}{\alpha I\left(\delta^h\right)_1} d\delta^h = \frac{-(Wb+a)}{Z^h q} d\delta^h \Longrightarrow \frac{dq}{d\delta^h} = \frac{-2\left(\frac{W}{V}+t_p\right)}{Z^h} \tag{37}$$

Equating the expression (37) to that obtained by differentiating the Eq. (36) with respect to the distance, we have:

$$\frac{\partial\left[q^h\left(\delta^h\right)\right]}{\partial\delta^h} = \frac{-2\left(\frac{W}{V}+t_p\right)}{Z^h\left(\delta^h\right)} = -2\left(\frac{t_p q_0}{®}\right) e^{-2t_p\left(\delta^h/®\right)} \tag{38}$$

where ® can be considered as constant because, being $\alpha I\left(\delta^h\right)_3 \equiv ® = @ - 2\alpha(W/V)\delta^h$ where @ is a constant for a nominal interest rate given, we can also consider for this case the term $(W/V)\delta^h$ as a constant for simplicity without loss of rigor. Then, by solving the Eq. (38), we have the optimum distance to the center for this class of consumers:

$$\delta_4^h = \frac{®}{2t_p} \ln\left[\frac{q_0 Z^h}{®\left(1+\frac{W}{Vt_p}\right)}\right] \tag{39}$$

This distance to central market is minor than the corresponding to the previous case (δ_3^h), but it does not minimize land expenses plus transport, something that did happen in the case of the consumers with restrictions of land and time to work.

4 Socially Efficient Tariffs for Passenger Transport

If we presuppose the existence of a public sector or Government in this model, we must think that it holds the possibility of Pareto efficiency in public transport prices. Private transport, as a complementary good of public transport, cannot have too different prices than public transport. Following Quinet [32], an efficient tariffication can be applied to the transport sector, but it will have to be adapted to the existence of non-tradeable goods as security and time used. In this case, the passenger tariff must maximize the consumer welfare, and its value will be identical to the marginal transport cost measured in opportunity cost.

Knowing that the total time of consumer to make his/her total expenses is $\underline{T} = \Omega^h + \mathbf{L}^h + (2/V)\delta^h$, the time spent in transport will be $(2/V)\delta^h$; then the monetary value of the opportunity cost to invest this time in doing this transport is $W(2/V)\delta^h$, where W is the average price of the production factors unused during the time of transport (opportunity cost).

On the other hand, from the consumer budget constraint we have that: $Y^h = P_c C^h + q^h(\delta^h)Z^h + M_d^h + 2t_p\delta^h$. That is, the total price payed by the consumer in transport expenses is: $2t_p\delta^h$. The relevant price for the purpose of cost-benefit is the price that reflects the opportunity cost. Therefore, the price that reflect the opportunity cost of transporting a passenger must meet: $W(2/V)\delta^h = 2t_p\delta^h$. Hence, the socially efficient unit rate for passenger transport must be:

$$t_p = \frac{W}{V} \tag{40}$$

Up to this point, we have considered the existence of two different land prices laws. However, to each distance there is an only unitary price of land, and the land's price in the center of the city must be similar for these two laws, because it is unique.

Equating the land's prices in these two laws defined in Eqs. (20) and (36), for the same distance to central market, we will have:

$$q^h = q_0 e^{\frac{-2(\frac{W}{V}+t_p)}{@}\delta^h} = q_0 e^{\frac{-2t_p}{®}\delta^h} \tag{41}$$

By assuming a socially efficient tariffs system ($W/V = t_p$), we can derive:

$$@/2 = ® \tag{42}$$

Therefore, under these conditions, the two land's price laws (20) and (36) are converted in one unique land's price law, which we can write as follows:

$$q^h = q_0 e^{\frac{-4t_p}{@}\delta^h} \tag{43}$$

Also, under this land prices law, the four optimal distances calculated in the above cases, δ_1^h, δ_2^h, δ_3^h and δ_4^h once the socially efficient tariffs are replaced, they will be transformed in only two optimal distances. In this manner, under socially efficient tariffs for passenger transport, we have that:

$$\delta_1^h = \delta_4^h = \frac{@}{4t_p}\log_e\left(\frac{q_0 Z^h}{@}\right) = \frac{@}{4t_p}\log_e\left(\frac{q_0}{q_1}\right) \tag{44}$$

and:

$$\delta_2^h = \delta_3^h = \frac{@}{4t_p}\log_e\left(\frac{2q_0 Z^h}{@}\right) = \frac{@}{4t_p}\log_e\left(\frac{q_0}{q_2}\right) \tag{45}$$

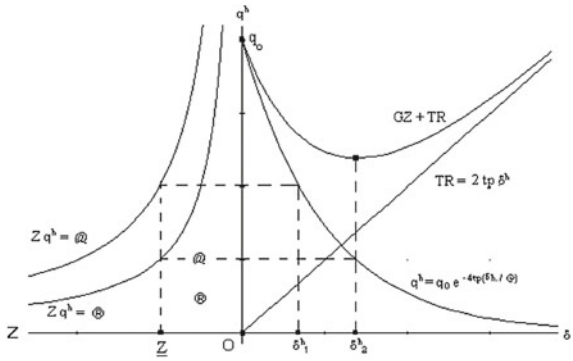


Fig. 1 Radial location of consumers around a Central Market (O) for land size: $\underline{Z}^h = 1$. 1) Distance in the general case (δ_1^h). 2) Distance when land size and time to work are fixed (δ_2^h). Land expenses (GZ); Transport expenses (TR)

being q_1 and q_2 the unitary land prices at distances δ_1^h and δ_2^h. The consumers who choose the distances δ_2^h and δ_3^h minimize land plus transport expenses (see Fig. 1). As can be seen by subtracting equation (44) from (45), the width of the circular crown where the consumers are located depends positively on consumer expenditure on land and negatively on the value of efficient rate of passenger transport in the city.

5 The Choice of Firms

The rational behavior of each producer is to maximize the benefit conditioned by a production function as follows: $\varphi^f(\mathbf{X}^f, \mathbf{L}^f, Z^f) = 0$, where $\mathbf{X}^f = (X_1^f, X_2^f, \ldots, X_n^f)$ and $\mathbf{L}^f = (L_1^f, L_2^f, \ldots, L_m^f)$. Land (Z^f) plays now the role of non-produced input. Urban concentration entails the existence of agglomeration economies that incorporate increasing returns to scale, as we see in Henderson [20], which could generate endogenous growth. The production function of the commodities (X_i^f) could then be expressed as follows:

$$X_i^f = \psi^f(H, F)\, f\left(Z^f, \mathbf{L}_j^f\right) = X_i^f\left(Z^f, \mathbf{L}_j^f\right) \tag{46}$$

In this function there are two types of inputs: Land (Z^f), whose price results differentiated by the distance to central market generating imperfect competition, and the other inputs used by firm (L_j^f), whose markets are assumed to be perfectly competitive. Following Sakashita [34], increasing returns are warranted by means of a certain function (ψ^f) that is an indicator of the agglomeration which depends on the consumers' number (H) and on the number of firms (F); it must fulfill the following requirements: $\psi'_H > 0$; $\psi''_{HH} < 0$; $\psi'_F > 0$; $\psi''_{FF} < 0$. Then, for each firm (f), the profit function will be:

$$B^f = \sum_{i=1}^{n}\left[P_i\left(X_i^f\right)X_i^f\right] - \sum_{j=1}^{m}\left(W_j L_j^f\right) - (H/F)\,t_m\delta^f - q^f\left(\delta^f, Z^f\right)Z^f \quad (47)$$

where $\sum_{i=1}^{n}\left[P_i\left(X_i^f\right)X_i^f\right]$ is the income coming from selling the X_i^f goods; the production costs are $\sum_{j=1}^{m}\left(W_j L_j^f\right)$ and the transport costs when final goods are transported from the factory to the central market are $(H/F)\,t_m\delta^f$. The fact that in the benefit function (47) the unit price of land q^f appears as dependent on the proper land is due to that while in the problem of the optimal consumer choice the prices of goods and income are given in the short term, in the profit maximization problem under imperfect competition the cost of using an input, such as land, is $q^f\left(\delta^f Z^f\right)Z^f$, because in this case the land price is not constant: it is a generic price on the land demand function as input. In both cases the unit price of the land depends on the distance of the consumer or producers from the center. The demand function of land as input is unknown but it must be derived from the producers' profit maximization. The producers' rational behaviour is to maximize profit according to a production function that guarantees technical efficiency:

$$\max\langle B^f = \mathbf{P}\left(\mathbf{X}^f\right)\mathbf{X}^f - \mathbf{W}\mathbf{L}^f - (H/F)\,t_m\delta^f - q^f\left(\delta^f, Z^f\right)Z^f\rangle \quad (48)$$

subject to: $\mathbf{X}^f = \mathbf{X}^f\left(Z^f, \mathbf{L}^f\right)$.

The Lagrangean function, in this case is:

$$\mathcal{L} = \mathbf{P}\left(\mathbf{X}^f\right)\mathbf{X}^f - \mathbf{W}\mathbf{L}^f - (H/F)\,t_m\delta^f - q^f\left(\delta^f, Z^f\right)Z^f + \mu^f\left[\mathbf{X}^f - \mathbf{X}^f\left(Z^f, \mathbf{L}_j^f\right)\right] \quad (49)$$

where μ^f is the Lagrange multiplier. Once solved this maximization we can deduce that the price paid by the producers for using land as an input (q^f), decreases when the distance to the central market increases:

$$\frac{\partial\left[q^f\left(\delta^f\right)\right]}{\partial\delta^f} = \frac{-Ht_m}{FZ^f} \quad (50)$$

On the equilibrium in land and transportation markets we assume that the unitary price of land shall be identical at same distance to CBD, for all economic agents. We know the land prices trajectory paid by the consumers in the area around the central market; but we do not know if this trajectory is the same to the producers. However, if we substitute the production function into the profit function, then expression (47) becomes in:

$$B^f = \sum_{i=1}^{n}\{P_i\left[X_i^f\left(Z^f, \mathbf{L}^f\right)\right]X_i^f\left(Z^f, \mathbf{L}^f\right)\} - \sum_{j=1}^{m}\left(W_j L_j^f\right) - (H/F)\,t_m\delta^f - q^f\left(\delta^f, Z^f\right)Z^f \quad (51)$$

And maximizing the profit function with respect to Z, we obtain:

$$B'_Z = \mathbf{P}'_Z \mathbf{X}^f + \mathbf{P}\mathbf{X}'^f_Z - \left(q'^f_Z Z^f + q^f\right) = 0 \tag{52}$$

rearranging this last expression, we have:

$$\mathbf{X}'^f_Z \left(\mathbf{P}'_X \mathbf{X}^f + \mathbf{P}\right) = \left(q'^f_Z Z^f + q^f\right) \tag{53}$$

But the term $(P'_{iX} X_i^f + P_i)$ is the marginal income corresponding to the goods X_i when good markets are in imperfect competition. Knowing that $Z^h q(\delta^h) = \alpha I\left(\delta^h\right)_1 = @$, considering the Marshallian demand functions coming from Eq. (8) when $\rho = \alpha$ the expenses in the composite good C^h is $P_c C^h = @$, and hence the demand equation of the X_i goods will be $X_i = @/P_i$; this demand function is a rectangular hyperbola and its corresponding marginal income is always zero. Hence, we can obtain:

$$q'^f_Z Z^f + q^f = 0 \tag{54}$$

But $(q'^f_Z Z^f + q^f)$ is the marginal income corresponding to the land demand function (Z^f) as input, whose market is in imperfect competition. Hence, we can express the demand function of land as input, as follows:

$$Z^f = ©/q^f \tag{55}$$

where in this case (©) is the value of the producer's expenditures in land as input, being q^f the unitary price of land. Knowing the form of the demand function of land as an input, we can establish some relations between land and transportation markets. If we substitute the result (55) in the Eq. (50), we have:

$$\frac{d(q^f)}{q^f} = \frac{-Ht_m}{F©} d(\delta^f) \tag{56}$$

By integrating (56), considering © as a constant:

$$q^f = q_0 e^{\frac{-Ht_m}{F©}\delta^f} \tag{57}$$

this is the relationship between land prices and distance to the central market, from the point of view of producers. Moreover, central land's price q_0 should be the same price in the consumer problem than in the producer problem, and hence under socially efficient tariffs for passenger transport q_0 must be the same in the formulations (43) and (57):

$$q_0 = q^h e^{\frac{4t_p}{@}\delta^h} = q^f e^{\frac{Ht_m}{F©}\delta^f} \tag{58}$$

This relation will also fulfill the equivalence between land prices and distances for both agents, consumers and producers. Hence, if $\delta^h = \delta^f$, then $q^h = q^f$. By

identifying coefficients in expression (58) we can obtain the value of the expenses that each firm make in land as input (©):

$$© = \left(\frac{Ht_m}{4Ft_p}\right) @ \tag{59}$$

where as is known, @ is the expenditure that the consumer makes on land as a good.

By introducing the expression (59) into the expression (57), and taking logarithms, we can know the optimal distance between the producer and the central market:

$$\delta^f = \frac{F©}{Ht_m} \log_e \left(\frac{q_0 Z^f}{©}\right) = \frac{@}{4t_p} \log_e \left[\left(\frac{4Ft_p}{Ht_m}\right)\left(\frac{q_0 Z^f}{@}\right)\right] \tag{60}$$

Seeing this formulation, the question is how δ^f can be related to the distances δ_1^h or δ_4^h, and δ_2^h or δ_3^h. That is, what is the relationship between the distance from the firms to the central market and the two different distances from the consumers to the central market. On the other hand, if we minimize respect to δ^f the total spending on transportation and land for each producer, subject to the land prices law of the present problem: $\min_{\delta^h} \langle (H/F)\, t_m \delta^f + q^f \left(\delta^f, Z^f\right) Z^f \rangle$ subject to $q^f = q_0 e^{\frac{-Ht_m}{F©}\delta^f}$, we have the same result than in Eq. (60), in this case, coming from the profit maximization. That is, it is not equal that in the consumer's case. An important difference between consumer and producer cases is that we have assumed that the consumer may have land restrictions while firms do not have them, which may determine a different location. To the producers the optimum distance to CBD coming from profit maximization is the same as minimizing land and transport costs.

6 Market Areas Distribution

The most important element of a metropolitan area is the central or main town which has a market area with the highest density of population of the metropolitan area; this is the city where all goods are traded. If it is a big city is likely to function as a monocentric city. Around this central market area there are other market areas with less population density, where some secondary products are marketed. We assume that all market areas have a circular shape around their markets and have the same surface. If the demand of consumption goods must be supplied throughout the space, then the central market area should be fully surrounded by other circular areas tangents as secondary markets. There are two ways to cover the entire space with circular areas: a square distribution of market areas or a triangular distribution. The square distribution embodies a loss of surface area covered 5.3 times that the triangular distribution. Therefore triangular distribution is more efficient than the square. Triangular expansion generates a hexagonal arrangement of the market areas enclosing a central area where generally a central market is. In the absence of central markets

market areas space can also be developed in rectangular or parallelogram forms. Due to the high population density around a metropolitan area and hence the scarcity of land, it is very likely that a metropolitan area develop its market areas approximately in an hexagonal form. This form will tend to a hexagonal regular form in absence of geographical roughness or proximity to a coast line, as mentioned in Christaller [7], Lösch [27], Beckmann [4] or Parr [30] central place theory approach. Contiguous to the basic hexagonal polygon which determine the central market area there are other six secondary market areas. The centers of these secondary market areas into the metropolitan area form other big hexagonal polygon, they have not identical populations and normally do not fulfill the Zipf [42] law, as occur for example, in the metropolitan area around Madrid, Spain, which is hexagonally developed. In the case of Madrid metropolitan area, these sub-centers are: San Sebastian de los Reyes, Torrejon, Arganda, Parla, Mostoles and Las Rozas.

Moreover beyond the metropolitan area, the hexagonal network can follow extending if the central place has a gravitational force sufficiently powerful concerning to the economic activity, trade, population or generation-attraction of travels, as there are no geographical obstacles.

From the equations (44), (45) and (60), let us now see how much land size (Z_1^f) must be leased by firms that are located at the same distance from the center as consumers who have no working time or land restrictions ($\delta_1^f = \delta_1^h$):

$$\delta_1^h = \frac{@}{4t_p} \log_e \left(\frac{q_0 Z^h}{@} \right) = \frac{@}{4t_p} \log_e \left(\frac{q_0 Z_1^f}{©} \right) = \delta_1^f \tag{61}$$

so, the relationship between land sizes will be:

$$\frac{Z^h}{@} = \frac{Z_1^f}{©} \tag{62}$$

In the same way, when producers are located at the same distance from the center as consumers with land size and working time restrictions ($\delta_2^f = \delta_2^h$), we have:

$$\delta_2^h = \frac{@}{4t_p} \log_e \left(\frac{2q_0 Z^h}{@} \right) = \frac{@}{4t_p} \log_e \left(\frac{q_0 Z_2^f}{©} \right) = \delta_2^f \tag{63}$$

We can observe then that at distance $\delta_1^f = \delta_1^h$ the land size for the firm is $Z_1^f = (©/@)Z^h$, while at the distance $\delta_2^f = \delta_2^h$ it will be $Z_2^f = 2(©/@)Z^h$, i.e. twice: $Z_2^f = 2Z_1^f$. At the distance δ_2^h from the center producers and the consumers who minimize land and transportation costs when land and labour time are fixed, could coincide. These consumers and the producers will be situated over a circumference with radius $\delta^f = \delta_2^h$ around the central market. Moreover, consumers and producers will not be mixed along these circumferences because generally firms cause negative externalities on consumers, whose would escape if possible away from the proximity of the firms. Besides that, from the supply side, the concentration of this firms causes

increasing returns to scale; these two reasons cause that firms will be concentrated in a few points along both circumferences with radius δ_1^f and δ_2^h.

We will see now what are the value of the land prices q_1 and q_2 corresponding to the distances $\delta_1^f = \delta_1^h = \delta_1$ and $\delta_2^f = \delta_2^h = \delta_2$, regarding q_0. By dividing the expression (45) by (44), we have:

$$q_1 = 2q_2 \tag{64}$$

That is, land price at distance δ_1 from central market is double than at distance δ_2. Regarding the expressions (44) and (45) for the consumers' location, the difference between the efficient distance when land and labour time are fixed for consumers (δ_2) and the optimum distance (δ_1), under an efficient system of passenger tariffs is:

$$\delta_2 - \delta_1 = \frac{@}{4t_p} \log_e 2 \tag{65}$$

These two distances form two circumferences with radii δ_1 and δ_2 around the central market, and the part of the population with labour time and land size fixed, who minimizes land and transportation costs, will be located over the circumference of radius δ_2.

Other relevant question is to determine the value of the unitary land price at the distance δ_1, that is q_1, relative to the land price al central market q_0. For that we will analyze the slopes m_1 and m_2 of the land prices law between the central market and the distances δ_1 and δ_2, after replacing the results (64) and (65):

$$\frac{|m_2|}{|m_1|} = \frac{\dfrac{q_1 - q_2}{\delta_2 - \delta_1}}{\dfrac{q_0 - q_1}{\delta_1}} = e^{\frac{4t_p}{@}(\delta_2 - \delta_1)} = \frac{1}{2} \tag{66}$$

Replacing Eq. (65) into (66) we have that: $\delta_1 = \frac{@}{4t_p} \log_e 2^{\left(\frac{q_0}{q_1} - 1\right)}$, and considering that Eq. (44) can be written as $\delta_1 = \frac{@}{4t_p} \log_e \left(\frac{q_0}{q_1}\right)$, equating these two relationships leads to the following expression:

$$\left(\frac{q_0}{q_1}\right) = 2^{\left(\frac{q_0}{q_1} - 1\right)} \tag{67}$$

which solutions for $(q_0/q_1) > 0$ are $(q_0/q_1) = 1$, and $(q_0/q_1) = 2$. But $(q_0/q_1) = 1$ is not a valid solution because $q_0 > q_1$ since the unitary land price in central market (q_0) must be more high than at distance δ_1. Hence:

$$q_0 = 2q_1 \tag{68}$$

Regarding from Eqs. (68) and (64) that $q_0 = 2q_1$ and $q_1 = 2q_2$, we have that $q_0 = 4q_2$. By replacing these results in Eqs. (44) and (45) we obtain that:

$$\delta_2^h = 2\delta_1^h \tag{69}$$

And following Eq. (65) these two distances are:

$$\delta_1 = \frac{@}{4t_p} \log_e 2 \tag{70}$$

$$\delta_2 = \frac{@}{4t_p} \log_e 4 \tag{71}$$

Equations (64), (69), (70) and (71) determine two concentric circles of radii δ_1 and δ_2 centered in the central market. These two circles have the property that an equilateral triangle can be circumscribed to the circumference of radius δ_1 while being inscribed to the circle of radius δ_2. This only happens under socially efficient rates in passenger transport. Consumers are not distributed evenly over the two circumferences, but are grouped in several villages spread over them. Consumers with restrictions in land size and in working time are located at the vertices of the triangle, whose distance to the center is δ_2, while consumers who maximize their utility subject to the budget and time constraints are located in the center of the sides triangle, whose distance to center is δ_1.

That is, consumers would be located in three nucleus on the circumference with radius δ_2 and in three other nucleus on the circle of radius δ_1. This triangular arrangement of space ensures the minimum linear distance between the two types of consumers, because the equilateral triangle is the regular polygon with less perimeter. Following the minimum distance principle, along each circumference consumers will be concentrated in three nucleus forming an equilateral triangle because of the externalities caused by firms. Following the principles of minimum distance among producers and maximum distance among producers and consumers due to industrial externalities, the producers will tend to be situated between each two nucleus of firms, forming other three nucleus of industrial concentration with its corresponding class neighborhoods, composing other equilateral triangle opposite to the triangle formed by the consumers properly. The six nucleus build a basic hexagonal regular polygon that will be reproduced through all metropolitan area.

To can develop a sequence based in regular hexagonal forms of market areas, the optimal unit price socially efficient for passenger transport must reach the following value, once considered the expressions (70) and (71):

$$t_p = \frac{@ \log_e 2}{4\delta_1} = \frac{@ \log_e 2}{2\delta_2} = \frac{W}{V} \tag{72}$$

which can be compared to the result of Eq. (40).

The firms can also be located beyond the distance δ_2, depending on the amount of land size needed. The expansion of the hexagonal market areas provides new points

Table 1 Land sizes and unit land prices for firms according to their distances to CBD

Distance to CBD	Land size	Unit land prices
$\delta_1^f = \delta_1$	$Z_1^f = Z_1^f$	$q_1 = q_0/2$
$\delta_2^f = 2\delta_1$	$Z_2^f = 2Z_1^f$	$q_2 = q_0/4$
$\delta_3^f = 2\delta_2$	$Z_3^f = 8Z_1^f$	$q_3 = q_0/16$
$\delta_4^f = \sqrt{7}\delta_2$	$Z_4^f = 2^{2\sqrt{7}-1}Z_1^f = 19.58Z_1^f$	$q_4 = q_0/39.16$
$\delta_5^f = \sqrt{13}\delta_2$	$Z_5^f = 2^{2\sqrt{13}-1}Z_1^f = 74.08Z_1^f$	$q_5 = q_0/148.16$
$\delta_6^f = 4\delta_2$	$Z_6^f = 2^7 Z_1^f = 128Z_1^f$	$q_6 = q_0/256$
$\delta_7^f = \sqrt{19}\delta_2$	$Z_7^f = 2^{2\sqrt{19}-1}Z_1^f = 210.51Z_1^f$	$q_7 = q_0/421.02$
$\delta_8^f = 5\delta_2$	$Z_8^f = 2^9 Z_1^f = 512Z_1^f$	$q_8 = q_0/1024$
$\delta_9^f = 2\sqrt{7}\delta_2$	$Z_9^f = 2^{4\sqrt{7}-1}Z_1^f = 776.95Z_1^f$	$q_9 = q_0/1553.9$
$\delta_{10}^f = \sqrt{31}\delta_2$	$Z_{10}^f = 2^{2\sqrt{31}-1}Z_1^f = 1124.85Z_1^f$	$q_{10} = q_0/2249.7$

Source Own elaboration

of space where unit land prices are minimal (Fig. 2). So, the next point belong to an adjoining hexagon and it will be at a distance $\delta_3^f = 2\delta_2$; considering equations (70) and (62), we have: $\delta_1 = \frac{@}{4t_p}\log_e 2 = \frac{@}{4t_p}\log_e\left(\frac{q_0 Z_1^f}{©}\right) = \delta_1^f$. Therefore:

$$2 = \frac{q_0 Z_1^f}{©} \tag{73}$$

In the same form: $\delta_3 = \frac{2@}{4t_p}\log_e 4 = \frac{@}{4t_p}\log_e\left(\frac{q_0 Z_3^f}{©}\right) = \delta_3^f$. Hence:

$$16 = \frac{q_0 Z_3^f}{©} \tag{74}$$

By dividing equation (74) into (73) we have that $Z_3^f = 8Z_1^f$. The next point of the hexagonal network is located at the distance $\delta_4 = \delta_2\sqrt{7}$ from central market: $\delta_4 = \frac{@}{4t_p}\log_e 4^{\sqrt{7}} = \frac{@}{4t_p}\log_e\left(\frac{q_0 Z_4^f}{©}\right)$. By dividing this expression into Eq. (73), we have: $Z_4^f = (4^{\sqrt{7}}/2)Z_1^f = 19.58Z_1^f$.

By reiterating this process of calculus for the different vertices of the hexagonal network we can determine the relationship between some possible distances from firms to the central market and its corresponding land sizes and unit land prices, which are collected in Table 1 and Fig. 2.

That is, under efficient passenger tariffs, if $\delta^f = \sqrt{3n+1}\delta_2$, then $Z^f = 2^{2\sqrt{3n+1}-1} Z_1^f$, where n must be a positive and entire real number. These last two laws are valid in a regular hexagonal network if the ratio $(n-3)/4$ is a positive, but non entire real number. Conversely, if a firm needs to have a land size bigger than Z_1^f then the ratio Z^f/Z_1^f must be equal to $2^{2\sqrt{3n+1}-1}$ and the firm must be located at the distance $\delta^f = \sqrt{3n+1}\delta_2$ from the central market. We can also extract the value of the unit land prices at different distances from CBD as in Table 1 and Fig. 2. Under these

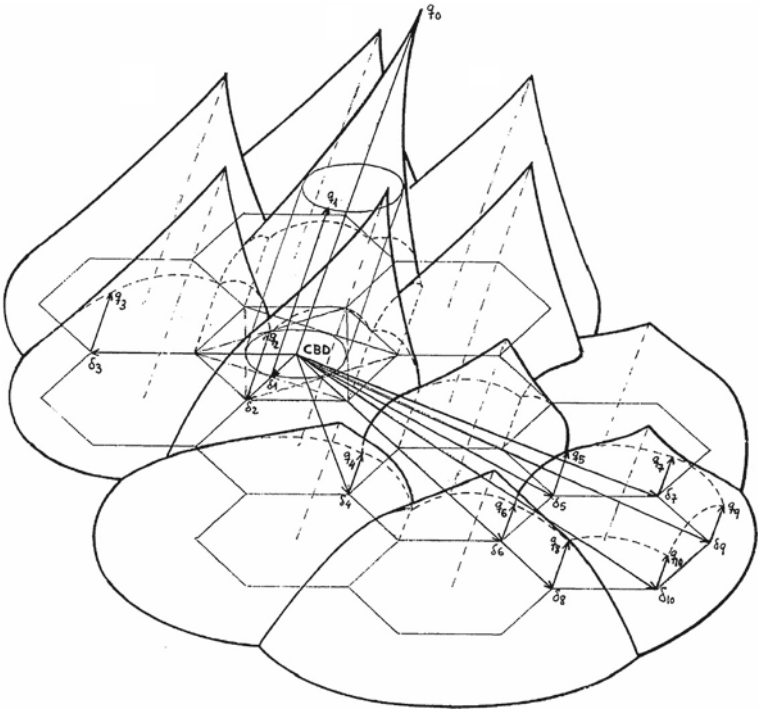


Fig. 2 Unit land prices distribution around the main city in a metropolitan area extended in hexagonal form. Beyond δ_2 firms are subject to a land prices law that collect the minimum unit land prices corresponding to the intersections of the price cones of the sub-center cities

circumstances, if the cities system is an hexagonal regular system and by assuming that each firm needs generally more land size than each optimizer consumer, firms will be generally located farther than consumers with respect to the central market. Finally, following Barreiro–Pereira [3], it can be shown that the total number of equations of the model: $(n + m + 4)(H + F + 1) + H + 2$, is equal to the number of unknowns.

7 Conclusions

This work represents a certain extension of the Herbert-Stevens and Alonso-Muth location models, by using Cobb-Douglas utility functions that generate unit elasticity demand functions. The aim of this work is to achieve the rational and efficient location of consumers and producers around a big central place, through a non-competitive general equilibrium microeconomic model where socially efficient fares for passenger transport are assumed. The difference with the models of the new economic

geography is that the present model emphasizes the role of the central place concept, crucial to explain the behaviour of large cities, trying to circumvent the importance of competition among cities. Unlike Eaton and Lipsey [8], in the present research firms have increasing returns to scale and occupy space. The present research explain the location of consumers and producers with respect to the center of a central place, where we assume that the central market is located. Agents are rational and efficient, that is, consumers maximize their utility conditioned by two constraints: a budget and a temporary one, while producers maximize profit. The model is closed by the Baumol-Tobin rule for transaction money demand, which implies that the nominal interest rate affects the location of consumers and firms in relation to the central market. The theoretical conclusions of the model indicate that the optimum distance to which a rational consumer is located with respect to the center tends to be greater the greater unit land price in the center, the amount of land acquired and the speed of passenger transport in the city; and the distance tends to be smaller the higher the consumers wage and the passenger transport rates. More results indicate that consumers are located at a distance from the central market that depends on whether the land size of their home and their working time are fixed or not. Only the consumers with time to work exogenously fixed and restrictions in their land size will minimize jointly their expenses in land and transport costs. They will be located farther than the consumers without restrictions, with respect to the CBD. When both situations occur, the consumers will be placed inside a circular crown with center in the central market, whose radii are the two distances to the center of the two types of consumers. The width of the circular crown where the consumers are located depends positively on consumer expenditure on land and negatively on the value of efficient rate of passenger transport in the city. If this rate is less than the socially efficient then the city expands. With respect to the optimal distance from the producers to the central market it results directly proportional to the number of firms, to the unit price of land in the CBD and to the amount of land purchased; and it results inversely proportional to the total number of consumers, and to the freight transport rate. Other results of the model indicate that the optimal amount of land purchased by the consumer is increasing with their incomes and nominal wages, but it is inversely proportional to the unit land price in the center and the nominal interest rate. The expenditure on land of consumers and producers also depends inversely on the nominal interest rate and therefore also on the expected rate of inflation. This makes the optimal distance of consumers and producers to the city center tends to be inversely dependent on the nominal interest rate and the expected inflation rate. Finally, it has also been shown in this work that the relationship between freight and passenger tariffs should not be arbitrarily established, because they should be related to the ratio between the consumption of land by producers and consumers, all of which could be taken into account by urban planners.

References

1. Alonso, W.: A reformulation of classical location theory and its relations to rent theory. Pap. Reg. Sci. **19**, 23–44 (1967)
2. Baumol, W.J.: The transaction demand for cash: an inventory-theoretic approach. Q. J. Econ. **66**, 545–556 (1952)
3. Barreiro-Pereira, F.: Determinantes Espaciales del Equilibrio Economico. Comunidad de Madrid, Consejeria de Economia e Innovacion Tecnologica, Madrid (2005)
4. Beckmann, M.J.: City hierarchies and the distribution of city sizes. Econ. Dev. Cult. Change **6**, 243–248 (1958)
5. Chamberlin, E.: The Theory of Monopolistic, Competition edn. Harvard University Press, Cambridge, Mass (1933)
6. Chen, A., Partridge, M.: When are cities engines of growth in China? Spread and backwash effects across the urban hierarchy. Reg. Stud. **47**, 1313–1331 (2013)
7. Christaller, W.: Die Zentralen Orte in Süddeutschland. Gustav Fisher Verlag, Jena (1933). Translated to english: Central places in Southern Germany, Prentice-Hall, Englewood Cliffs, N.J. (1966), by C.W.Baskin
8. Eaton, B.C., Lipsey, R.G.: An economic theory of central places. Econ. J. **92**, 56–72 (1982)
9. Fisher, I.: The Purchasing Power of Money. Macmillan, New York (1911)
10. Fujita, M.: Urban Economic Theory: Land Use and City Size. Cambridge University Press, Cambrigde, Mass (1989)
11. Fujita, M., Krugman, P.: A monopolistic competition model of urban systems and trade. Department of Regional Science, University of Pennsilvania, Mimeo, Philadelphia, PA (1992)
12. Fujita, M.: Monopolistic competition and urban systems. Eur. Econ. Rev. **37**, 308–315 (1993)
13. Fujita, M., Krugman, P.: When is the economy monocentric? von Thnen and Chamberlin unified. Reg. Sci. Urban Econ. **25**, 505–528 (1995)
14. Fujita, M., Mori, T.: Structural stability and evolution of urban systems. Reg. Sci. Urban Econ. **27**, 399–442 (1997)
15. Fujita, M., Krugman, P., Mori, T.: On the evolution of hierarchical urban systems. Eur. Econ. Rev. **43**, 209–251 (1999)
16. Fujita, M., Krugman, P., Venables, A.: The Spatial Economy. The MIT Press, Cambridge, Mass (1999)
17. Glaeser, E.: The Economic Approach to Cities. Harvard University, John F. Kennedy School of Government, Research Working Paper Series, RWP08-003 (2008)
18. Greenhut, M.L., Lee, C., Norman, G.: Product differentiation and intensity and sensitivity. In: Spatial Microeconomics, Edward Elgar, Aldersot, UK (1995)
19. Hart, O.: A model of imperfect competition with keynesian features. Q. J. Econ. **97**, 109–138 (1982)
20. Henderson, V.: The effects of urban concentration on economic growth. In: NBER, Working Paper Series, WP7503 (2000)
21. Herbert, J.D., Stevens, B.H.: A model of the distribution of residential activity in urban areas. J. Reg. Sci. **2**, 21–36 (1960)
22. Hsu, W.-T.: Central place theory and city size distribution. Econ. J. **122**, 903–932 (2012)
23. Hsu, W.-T., Holmes, T.: Optimal city hirarchy: a dynamic programming approach to central place theory. J. Econ. Theory **154**, 245–273 (2014)
24. Isard, W.: General equilibrium of the economic subsystem in a multiregional setting. In: Location Analysis and General Theory. Macmillan, 1st v. MIT Press, Cambridge, Mass (1990)
25. Kuenne, R.: General Equilibrium Economics: Space, Time and Money. Macmillan, Hong Kong (1992)
26. Krugman, P.: Geography and Trade. MIT Press, Cambridge, Mass (1991)
27. Lösch, A.: Die Räumliche Ordnung der Wirtschaft (1940). M. Fisher, Jena. Translated to english: The Economics of Location, New Haven, Conn., Yale University Press. (1954), by Woglom and Stolper

28. Mori, T.: A modeling of megalopolis formation: the maturing of city systems. J. Urb. Econ. **42**, 133–157 (1997)
29. Muth, R.F.: The derived demand for urban residential land. Urban Stud. **8**, 243–254 (1969)
30. Parr, J.B.: The law of market areas and the size distribution of urban centers. Pap. Reg. Sci. **76**, 43–78 (1997)
31. Pred, A.: City Systems in Advanced Economies: Past Growth, Present Process, and Future Development, Options edn. Wiley, New York (1977)
32. Quinet, E.: Transports et Théorie Économique. Presses de l' École Nationale de Ponts et Chaussées, Paris (1992)
33. Quinzii, M., Thisse, J.-F.: On the optimality of central places. Econometrica **58**, 1101–1119 (1990)
34. Sakashita, N.: Optimal utilization of the CBD with economy and diseconomy of agglomeration. In: Ohta, Thisse (eds.), Space, Does Economic, Matter? S.Martin's Press, London (1993)
35. Sassen, S.: Cities In a World Economy. Pine Forge Press, Thousand Oaks, California (2000)
36. Small, K.A.: Urban Transportation Economics. Harwood Academic Publishers GmbH, Luxembourg (1992)
37. Stahl, K.: Location and spatial pricing theory with nonconvex transportation cost schedules. Bell J. Econ. **13**, 575–582 (1982)
38. Tabuchi, T., Thisse, J.-F.: A new economic geography model of central places. J. Urban Econ. **69**, 240–252 (2011)
39. Thisse, J.F.: Oligopoly and the polarization of space. Eur. Econ. Rev. **37**, 299–307 (1993)
40. Tobin, J.: The interest elasticity of transactions demand for cash. Rev. Econ. Stat. **38**, 241–247 (1956)
41. von Thünen, J.H.: Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalkonomie, Perthes, Hamburg. Translated to english: von Tünen's Isolated State, Pergamon Press, Oxford. (1966), by C.M. Wartenberg
42. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Cambridge, Mass (1949)

Predicting Energy Demand in Spain and Compliance with the Greenhouse Gas Emissions Agreements

Diego J. Bodas-Sagi and José M. Labeaga

Abstract This paper aims to predict energy demand in Spain for the year 2020 and analyzes whether this country will be able to meet the European Union's greenhouse gas emission reduction commitment. To this purpose, we use climatic data and some variables to measure the economic activity in Spain. The simulated scenario considers that Spain will begin a process of economic recovery which will result in an increase in industrial activity with stable climatic conditions. Several techniques including Simple Linear Regression, Support Vector Machines or Deep Learning have been proposed to estimate and test the model. The EU agreements imply that by 2020 between 20 and 30% of the consumed energy will come from clean and renewable energy sources. The conclusions for this paper show that Spain may be on track to meet its commitments to Europe.

Keywords Energy demand · Greenhouse gas emission · Linear models

1 Introduction

This paper uses a very simple model to predict energy demand in Spain for the 2020 scenario. We based our prediction solely on climatic and some variables proxying the economic activity. Our main research question is whether Spain can achieve an energy-mix able to meet the EU commitments by 2020. In general, forecasting the likely path of greenhouse gas emissions is essential to understanding the range of possible effects of climate change. The European Commission (EC) and EU governments agreed on the target of cutting greenhouse gases by at least 20% by 2020 [1], compared with 1990 levels. Hence, it is mandatory for all EU member

D. J. Bodas-Sagi (✉) · J. M. Labeaga
Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia,
Paseo Senda del Rey, 11, 28040 Madrid, Spain
e-mail: diegobodas@yahoo.es

J. M. Labeaga
e-mail: jlabeaga@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_5

countries to ensure that between 20 and 30% of the consumed energy comes from clean renewable energy sources. This European action on climate change has its antecedents in Articles 17, 18 and 19 of the Directive EU 2009/28/CE of the European Parliament and Council of April 23, 2009, which was transferred to Spain by the RD 1597/2011 of November 4, 2011. The United Nations Climate Change Conference (Paris, December 2015) ratified these agreements. They have been later approved by the European Union Parliament on October 4, 2016.

The methodology used for forecasting energy demand is manifold. Our model is parsimonious since we use a specification considering climatic variables (Heating Degree Days or HDD, Cooling Degree Days or CDD, and volume of rainfall), and economic variables (activity level proxied only by the Industrial Production Index or IPI).[1] The IPI measures the monthly development of industrial activity, including extractive, manufacturing, and production and distribution of electricity, water and gas. This indicator reflects the joint development of quantity and quality, independent of the influence of prices. The Instituto Nacional de Estadística (INE) builds the IPI through a survey concerning details of the production of activity branches compiling monthly data for more than 11,500 establishments. According to [2], the industrial sector is the largest consumer of electricity, close to 30% of the total amount.

A second objective of this paper is to compare different regression methods with prediction purposes. We use a Mean Square Error (MSE) criterion to test Linear Regression, Support Vector Machines for Regression (SVR) and Deep Learning Neural Networks. Deep Learning uses machine learning algorithms in order to model high-level abstractions with multiple non-linear transformations. In addition, for greater accuracy and sensitivity in the evaluation, we divide the original data into a training set and a test set, as is explained in Sect. 3. The results show that models based on neural networks significantly improve the MSE criterion when compared to Linear Regression or SVR. On the other hand, if we consider a scenario for the foreseeable future consisting of an increase of IPI similar to that given in previous reporting periods (from November 2008 to December 2011), the simulations predict an energy demand in December 2019 close to 23 thousand Gigawatt hours (GWh). According to our data, this demand will contribute to more than 6 million tons of CO_2 emissions to the atmosphere. Considering historical data from Red Eléctrica de España (REE) and the afore mentioned energy demand there is clear evidence that this energy-mix will allow Spain to meet EU clean renewable energy agreements and, that this will cover between 20 and 30% of total demand.

The remainder of the paper is organized as follows: Sect. 2 introduces and describes the data used for the empirical exercise; In Sect. 3, we explain the different methods and the results obtained; Sect. 4 shows the results of the simulation of energy demand at the end of 2019; Sect. 5 concludes.

[1] We try another economic indicators but since the industry is the largest consumer of energy, we believe our parsimonious model can fit better and cover our prediction purposes.

2 The Data

Our aim is to perform the analysis using a parsimonious model fed by as little information as possible. This study only uses climatic variables, and a proxy for industrial production. The model used has been tested against unrestricted models based on demand specifications and this is our preferred model for prediction purposes based on a battery of tests. According to a study published by the BBVA Foundation [3], industry, historically, transfers purchasing power to other sectors. Company profits evolve in line with a yearly exchange rate which is linked to the IPI. Economic crashes are likely to be reflected significantly in this index. The Spanish IPI (adjusted seasonally) reached its lowest value since 2007 in April 2012, with an accumulated depreciation close to 30% during this period [4]. Furthermore, annual series with mean monthly IPI values for the period 2007–2014, are highly correlated with average annual expenditure of Spanish households during that period. It is therefore assumed that the IPI provides an accurate proxy of the economic activity to be used for estimating different scenarios for economic growth in Spain.

In addition we use climatic variables. Temperature data have been obtained from several sources, including Agencia Estatal de Meteorología (AEMET) weather stations, Ministry of Agriculture, as well as data from Red Eléctrica Española (REE) and from the National Climatic Data Center [5]. Weather stations are located in different areas of Spain: the north; Cantabrian coast; the Meseta Central; the south and Mediterranean Coast. Matching this data and taking daily averages, we built a dataset with daily observations for the period March 1, 2007 to December 31, 2015, with estimated maximum, minimum and average daily temperature. We obtain daily rainfall (PREC) in the same way. Data from the islands (Balearic Islands and the Canary Islands), have not been considered in this paper. Based on these temperatures, HDD and CDD have been calculated using the following formulas [6], using the first one that matches:

$$HDD_t = \begin{cases} 0 & t_{min_t} > t_{base_t} \\ \frac{t_{base_t} - t_{min_t}}{4} & \frac{t_{max_t} + t_{min_t}}{2} > t_{base_t} \\ \frac{t_{base_t} - t_{min_t}}{2} - \frac{t_{max_t} - t_{base_t}}{4} & t_{max_t} \geq t_{base_t} \\ t_{base_t} - \frac{t_{max_t} + t_{min_t}}{2} & t_{max_t} < t_{base_t} \end{cases} \quad (1)$$

$$CDD_t = \begin{cases} 0 & t_{max_t} < t_{base_t} \\ \frac{t_{max_t} - t_{base_t}}{4} & \frac{t_{max_t} + t_{min_t}}{2} < t_{base_t} \\ \frac{t_{max_t} - t_{base_t}}{2} - \frac{t_{base_t} - t_{min_t}}{4} & t_{min_t} \leq t_{base_t} \\ \frac{t_{max_t} + t_{min_t}}{2} - t_{base_t} & t_{min_t} > t_{base_t} \end{cases} \quad (2)$$

where t represents the day. The results of Fig. 1 are taken as base temperature for calculating HDD and CDD the value of 15.5 °C. In order to match climatic data to economic data we take mean monthly values.

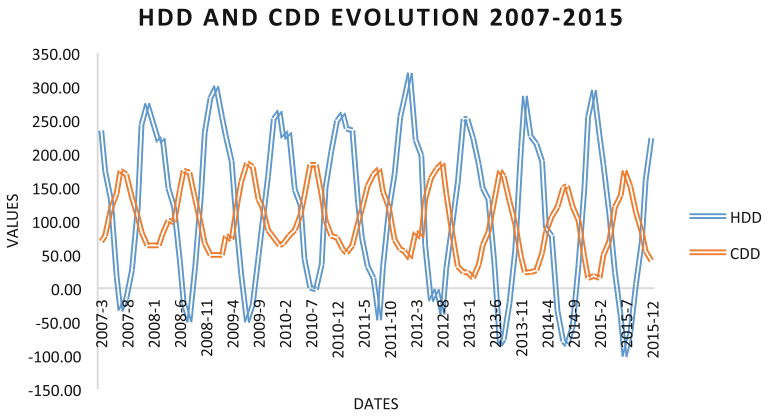


Fig. 1 HDD and CDD evolution over the period 2007–2015

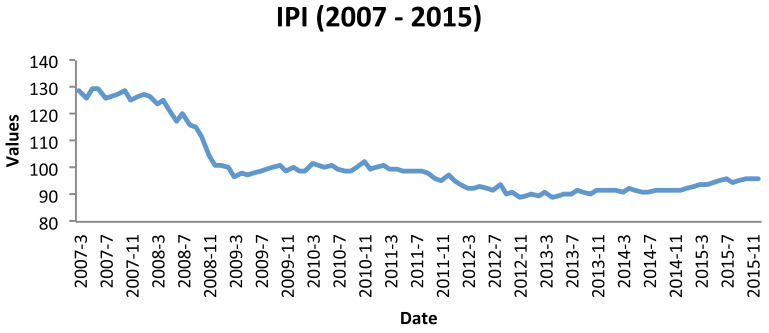


Fig. 2 IPI evolution for the period 2007–2015

We also need information about economic activity and prices. Several possibilities are available such as the Gross Domestic Product (GDP), unemployment data, energy prices, Consumer Price Index, etc. We have decided to use the Industrial Production Index (IPI), a monthly time series collected by INE. Available data are shown in Fig. 2.

For the purpose of comparison, quarterly GDP growth and unemployment data for the period considered in the analysis are shown in Fig. 3. This data are produced quarterly by the INE.

Figure 4 shows GDP adjusted taking into account unemployment.

The industrial sector is the largest consumer of electricity (30%), while the services sector accounts for 13% of consumption [2].

Finally, monthly energy demand data in GWh (from REE and the Ministry of Industry and Energy) are shown in Fig. 5.

As we observe all these figures it is difficult non-parametrically to decide upon the best proxy for energy demand. We can assume that a model which rationalizes the behavior of economic agents, and whose specification includes a proxy for income

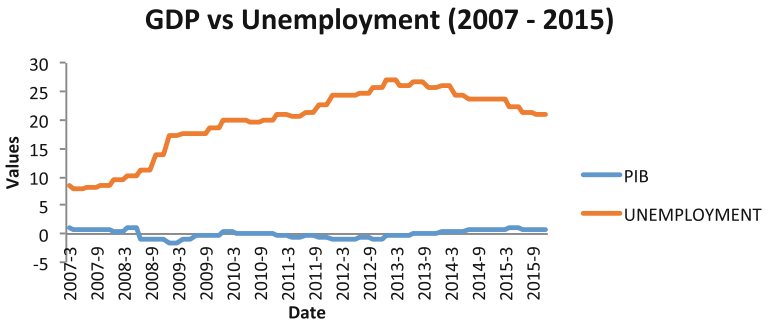


Fig. 3 GPD and unemployment evolution (2007–2015)

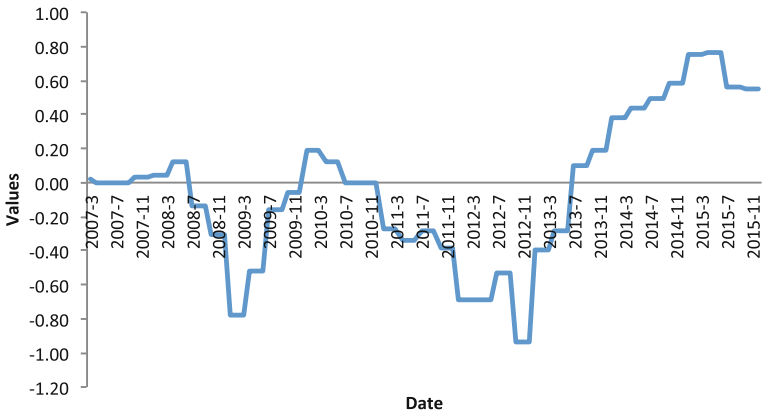
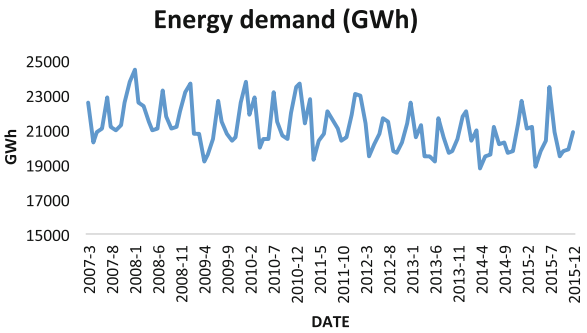


Fig. 4 GDP adjusted for unemployment (2007–2015)

Fig. 5 Energy demand in Spain



(GDP) and proxy for prices of energy could provide an adequate alternative, i.e., a proper model of demand. However, our main aim is to provide a parsimonious model to obtain the best possible prediction.

3 Methodology and Results

In this section we explain the different methods to adjust energy models, while also processing and showing the results. As mentioned previously, our goal is to predict energy demand (GWh.)[2] in Spain to test whether EU commitments can be achieved. The EU agreement forces countries to use between 20 and 30% of total energy using renewable or clean sources. We like to test whether parsimonious models help us in making accurate predictions of energy demand or energy consumption by only using climatic variables and the IPI index. Reference [7] shows that the demand for energy is absolutely inelastic with respect to the price for Spain in the considered period.

3.1 Estimation Methods

Taking into account the no free lunch theorems [8], we chose to evaluate different methods in order to disentangle the particular preferred technique according to the testing procedure. To be more precise, in this paper we use Linear Regression, Support Vector Machines and a Deep Learning algorithm. All these models are implemented using the R Software [9].

3.1.1 Linear Regression

Linear regression is a simple approach for predicting a quantitative response Y on the basis of a single regression variable X. It assumes that there is a linear relationship between X and Y [10]. We can write this linear relationship as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3}$$

where β_0 and β_1 are two parameters that represent the intercept and slope. ε represents the error and contains the variability of the dependent variable not explained by the X.

The regression coefficients β_0 and β_1 are unknown, and they are estimated on a sample ($\widehat{\beta_0}$ and $\widehat{\beta_1}$). With these estimated coefficients, we can obtain predictions ($\widehat{Y}$) as follows:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X \tag{4}$$

There are several approaches to obtain the parameters, one of most common approaches involves minimizing a Least Squares Criterion (LSC) [11]. Ordinary Least Squares (OLS) with heteroskedasticity-autocorrelation robust standard errors [12] has been selected for this work. If $\widehat{Y}$ is a vector of T predictions and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions, we choose $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize the residual sum of squares (RSS):

[2] We denote energy models as we cannot characterize them as demand or supply models. In any case, we acknowledge our interest in predicting energy consumption.

$$RSS = \sum_{1}^{T} (Y_i - \widehat{Y}_i)^2 \tag{5}$$

In practice, we often have more than one predictor, so Multiple Linear Regression (MLR) is used. Suppose that we have k different predictors, the MLR model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \tag{6}$$

In this case, k coefficients have to be estimated, let say $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\beta_k}$. In this case, we can obtain predictions ($\widehat{Y}$) as follows:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \widehat{\beta_3} X_3 + \ldots + \widehat{\beta_k} X_k \tag{7}$$

The values $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\beta_k}$ that minimize Eq. 5 are the multiple (ordinary in our case) least squares regression coefficient estimates [13].

3.1.2 Support Vector Machines

Support Vector Machines (SVM) has been widely used for function estimation [14, 15], known for our case Support Vector Regression (SVR). SVM is a generalization of a classifier called the maximal margin classifier [16] in order to accommodate non-linear class boundaries. SVM can be applied not only to classification problems but also to the case of regression. It contains all the main features that characterize maximum margin algorithm: a non-linear function is learned by linear learning machine mapping into a high dimensional kernel induced feature space and, the capacity of the system is controlled by parameters that do not depend on the dimensionality of the feature space. In SVR, the input X is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) $f(X, w)$ is given by:

$$f(X, w) = \sum_{1}^{k} w_i g_i(X) + b \tag{8}$$

where $g_i(X)$ and $i = 1 \ldots k$ denotes a set of nonlinear transformations. b is known as the bias term. The quality of estimation is measured by the loss function $L(y, f(X, w))$. SVR uses a type of loss function called ε insensitive loss function [14]:

$$L(y, f(X, w)) = \begin{cases} 0 & |y - f(X, w)| \leq \varepsilon \\ y - f(X, w)| - \varepsilon & otherwise \end{cases} \tag{9}$$

In addition to use an ε insensitive loss function, SVR tries to reduce model complexity by minimizing $||w||^2$ (see [17], for additional details). SVR has been adjusted using *e1071* R package [18]. In our work, we have use a linear kernel and epsilon support vector regression function.

By default, ε takes value of 0.1. But in order to improve the performance of the support vector regression we have executed a grid search looking for the best value for ε. There is also a cost parameter which we can change to avoid overfitting. The process of choosing these parameters is called hyperparameter optimization [19], or model selection. We do not have enough data to consider an extra validation set to caliber these parameters.

3.1.3 Deep Learning Neural Networks

Some techniques try to replicate the efficiency and robustness by which the human brain represents information and obtains knowledge. These works motivated the emergence of the subfield of deep machine learning, which focuses on computational models for information representation that exhibit similar characteristics to the neocortex [20]. Artificial Neural Networks (ANNs) are a family of deep learning models inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. ANNs are generally presented as systems of interconnected neurons that exchange messages among them. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. This technique has been extensively used for forecasting tasks with some good results [21]. As disadvantages we can quote its black box nature, its greater computational burden, its proneness to over-fitting, and the empirical nature of model development. A neural network can be thought of as a network of neurons organized in layers. The predictors or input form the bottom layer, and the forecasts or output form the top layer. Once we add and intermediate layer with hidden neurons, the neural network becomes non-linear. Configuring the neural network, activating function, layers, etc., are not trivial tasks.

3.2 Empirical Results

For accurate assessment, we randomly divide the original data into two sets (getting each set data from all the months), a training set and a test set. The test set contains data from January 2008 to June 2008 and from July 2014 to December 2014. The remaining data is used for training. We have randomly selected different months and years to insert some variability in the data.

Fig. 6 Results on test set using linear regression

As explained previously, our main objective is to estimate the parameters of the model represented by the following equation:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 IPI \tag{10}$$

The following figures display results graphically, in each case, we first show predicted values using test set, and subsequently, using training set. The following 2 figures refer to results using linear regression with ordinary least squares method (Figs. 6 and 7):

Using SVM, we obtain the following graphics (Figs. 8 and 9):

As we show in the next section, SVR error (MSE) exceeds that of the classical regression MSE.

In the case of Deep Learning Neural Networks, the architecture we have employed consists of two hidden layers with the same number of neurons. The neural network is trained with stochastic gradient descent using back-propagation. This architecture has been tested with 20, 50, 100 and 200 neurons in each layer. The results are shown Table 1. These results consider a base temperature value of 15.5.

Table 1 shows a known problem of over-fitting exhibited by this method, which may occur when using 100 and 200 neurons per layer. In these cases, the MSE

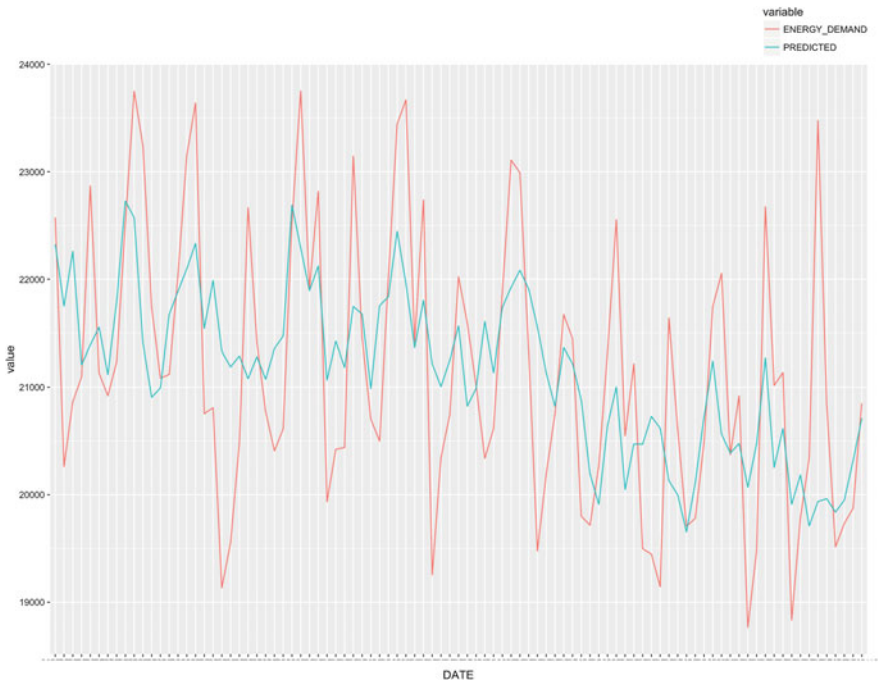


Fig. 7 Results on training set using linear regression

using the training set drops significantly compared with those obtained by other architectures. It becomes harder to avoid over-fitting with small-data and sometimes simpler models are more appropriate. However, the MSE using the test set is greater compared to those obtained using 50 neurons per layer. Figure 10 shows prediction results using the training set and 50 neurons in each layer (using a base temperature value of 15.5). The comparison of observed and predicted values in Fig. 10 indicates that the model adjusted through Deep Learning Neural Networks shows a rather high accuracy.

MSE using the training data is lower than using a simpler architecture, but it is significantly higher when the test sample is considered. Therefore, we will work with 50 neurons in each layer according to the tests reported in Table 1. Results using this architecture are shown in Fig. 11.

According to the graph, it may seem that this technique obtains a better approach. This feeling is confirmed by the comparison we will carry out next.

3.2.1 Comparison of Different Methods

To compare the results obtained by different techniques, we show in Table 2 the MSE on the test sample taking into account each technique and a base temperature.

Fig. 8 Results on test set using SVR

This comparison of the three procedures used points toward the use of deep learning results for predicting total energy demand. As we have previously explained, Linear Regression uses Ordinary Least Squares Criterion.

We can also show results using the coefficient of determination (R^2). An R^2 of 1 indicates that the regression line perfectly fits the data. Table 3 illustrates that the deep learning technique obtains the best value.

Coefficients for HDD, CDD, and intercept parameter are significant (p-value less than 0.001). Tests do not find significance for PREC (precipitations) or IPI coefficients. Despite the fact that industry sector is the largest consumer of electrical energy taking into account other productive sectors such as services (it has been mentioned in Sect. 2), models have not found significance for the IPI coefficients. We tried again using GDP (quarterly GDP growth in percentage) instead of IPI. In this case, and using linear model we found little significance for GDP coefficient (p-value between 0.1 and 0.05, and no find significance using other models). But we obtain worse results for MSE and R^2 in all cases and with all the techniques. For example, with a base temperature value of 14.5 the linear model using IPI obtains a MSE value in test of 973,182.1, but using the GDP as input variable instead of the IPI, the obtained value is 1,039,117. For this reason, we have chosen not to include GDP in this model.

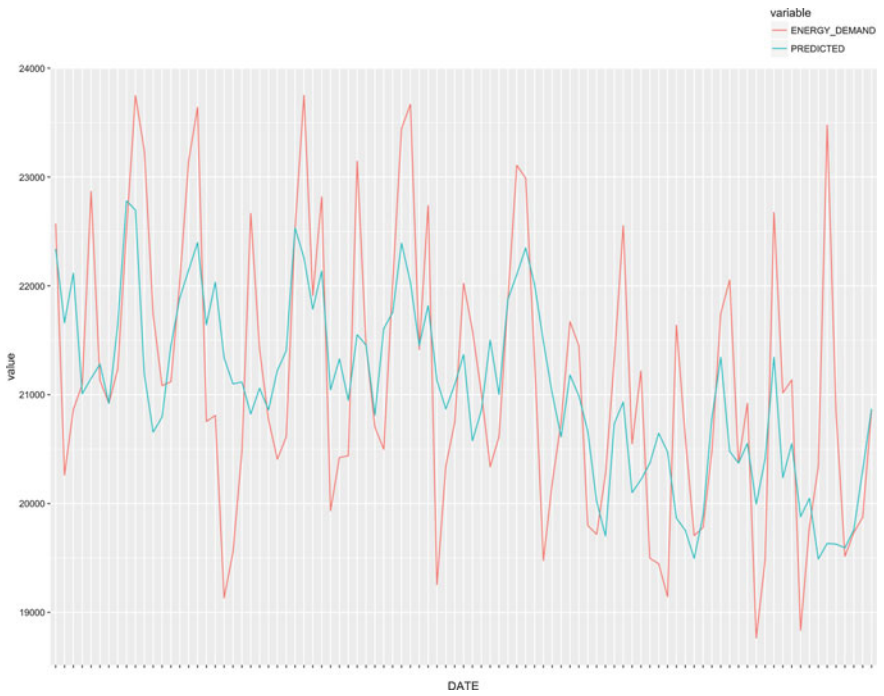


Fig. 9 Results on training set using SVR

Table 1 Deep Learning Neural Network experiments

Neurons in layer	MSE in test	MSE in training
20	1,168,439	256.4161
50	346,196.6	365.0929
100	385,315	125.3764
200	479,269.1	229.5746

Regarding the execution time in seconds required for each technique, the linear model obtains the results in 0.0017 s, the SVR lasts 55 s and deep learning methods take 2.8 min (on a laptop with 2,9 GHz Intel Core i5 and 8 GBs of RAM). We can conclude that for the problem at hand the lower values of the MSE criterion compensates for the longer execution time.

Tuning SVR parameters implies to increase execution time from 0.65 s to 46. Grid search sets the ε value to 0.1 using 14.5 as base temperature, while 0.8 and 0.7 have been the selected values for 15.5 and 16.5 base temperature values respectively.

Figure 12 shows results looking for ε best values, taking into account cost optimization and a base temperature value of 16.5.

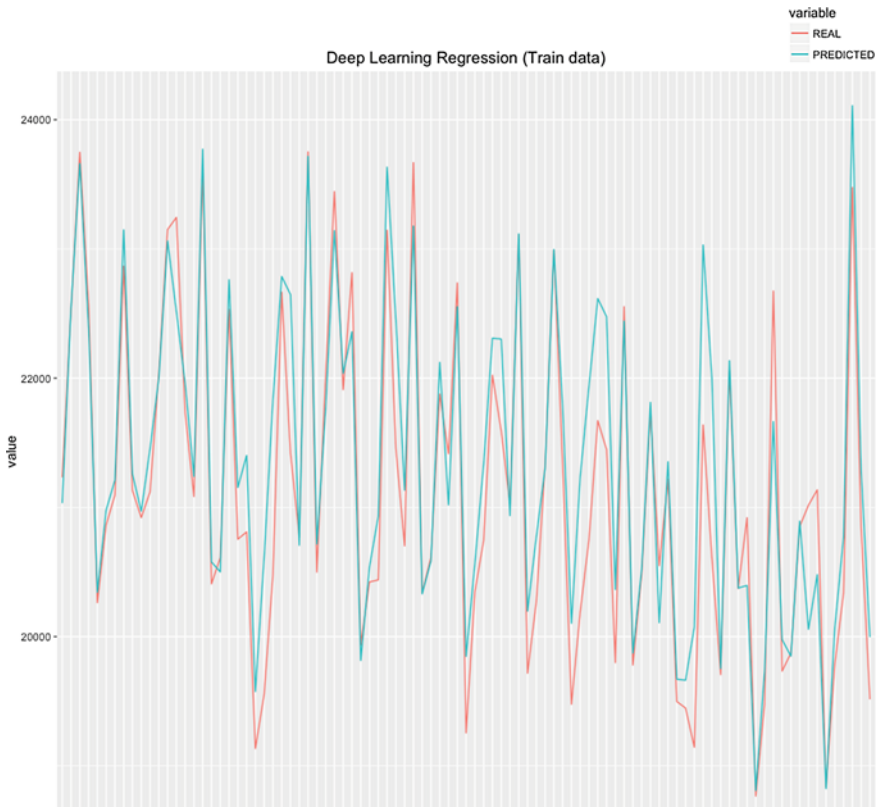


Fig. 10 Deep Learning Neural Networks applied to training set

In addition to comparing the different methods, we validate our model doing an analysis of variance (ANOVA) of monthly residuals. We do not detect evidence of heteroscedasticity (p-value is 0.19 greater than the reference value 0.01). However, when we are working with aggregated monthly data, the volume of information is insufficient for a given year. If we join residuals from all the time period, the ANOVA shows evidence of significant differences between the prediction errors generated for each technique.

3.3 Adjusting a Rational Demand Model

The previous model has not properly derived from any optimization problem taking into account behavior of economic agents. We postulate here a model that takes into account energy demand as a function of income and prices. (see, for instance, [22]). Some readers might find it strange that in the linear model, IPI coefficient shows

Fig. 11 Results on test set using Deep Learning Neural Networks

non statistical significance. In this context, we have tried other models to improve the MSE and R^2 results. The main difference is to consider the GDP deflator. The GDP deflator is a measure of price inflation/deflation with respect to a specific base year. Taking the nominal GDP (in millions of euros) and the GDP deflator we obtain the real GDP, we express it in logs. Now, to estimate energy demand we use the following equation:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) \quad (11)$$

The obtained MSE and R^2 values have been the following (Table 4).

We can also show results using the coefficient of determination (R^2) (Table 5).

Analyzing significance for coefficient values, we can conclude that coefficients for HDD, CDD, and intercept parameter are significant (p-value less than 0.001). Again, tests do not find significance for PREC (precipitations) or $\frac{nominalGDP}{GDPdeflator}$ coefficients.

Table 2 MSE of different methods

Base temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	973,182.1	1,027,007	511,557.6
15.5	970,816	1,017,013	346,196.6
16.5	968,461.8	1,007,333	358,933.9

Table 3 R^2 of different methods

Base temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.439	0.392	0.748
15.5	0.440	0.392	0.801
16.5	0.442	0.374	0.708

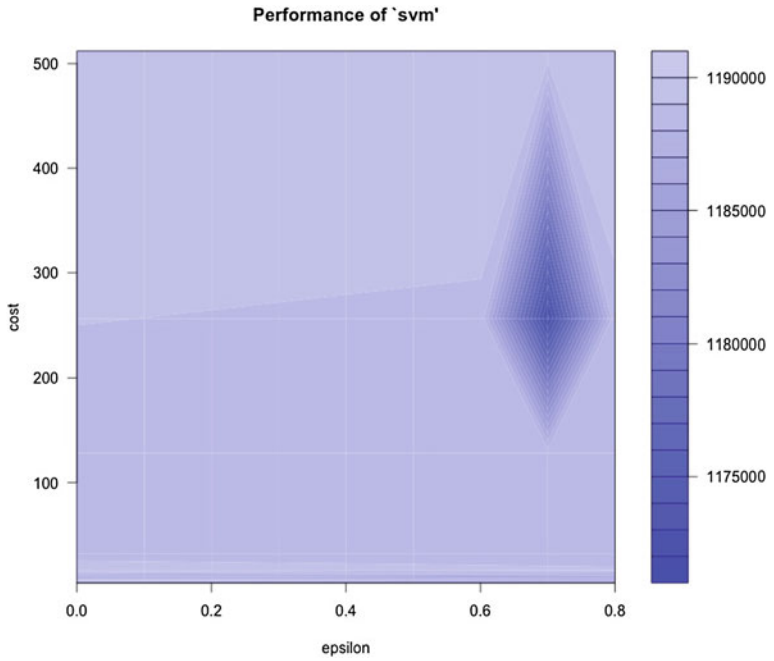


Fig. 12 Tuning ε value for SVR and base temperature 16.5

We have also evaluate results applying natural logs to energy demand as next equation shows:

$$log(ENERGY\_DEMAND) = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) \quad (12)$$

Table 4 MSE using nominal GDP and GDP deflator

Base temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,120,323	1,084,567	2,849,661
15.5	1,118,067	1,081,096	1,238,948
16.5	1,115,824	1,077,402	1,979,832

Table 5 R^2 using nominal GDP and GDP deflator

Base temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.35	0.37	Not applicable
15.5	0.36	0.38	0.29
16.5	0.36	0.38	Not applicable

Table 6 R^2 using nominal GDP and GDP deflator - taking logs in energy demand

Base temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.35	0.27	Not applicable
15.5	0.36	0.36	Not applicable
16.5	0.36	0.32	Not applicable

In this case, R^2 results (we cannot compare MSE values using different predicting values) are shown in Table 6.

Analyzing these results we can conclude that inserting in the equation the real GDP does not improve the predictive power of the model.

We can now add some extra information to the model taking into account the energy prices (EP - index with base = 100) and the CPI (monthly). The goal is to evaluate is this extra information is useful to improve the prediction. The Eq. 13 models this case:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) + \beta_5 log(\frac{EP}{CPI}) \quad (13)$$

Tables 7 and 8 show the results. This model does not improve the first model MSE or R^2.

Interpreting the results, we can conclude that coefficients for HDD, CDD, and intercept parameter are significants (p-value less than 0.001). We found slight significance for β_5 coefficient (p-value between 0.1 and 0.05. We don't find significance for PREC (precipitations) or $\frac{nominalGDP}{GDPdeflator}$ coefficients).

Table 7 MSE adding information about energy prices and CPI

Base temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,141,071	1,257,050	867,865.4
15.5	1,138,936	1,236,390	1,773,467
16.6	1,136,809	1,222,573	1,953,979

Table 8 R^2 adding information about energy prices and CPI

Base temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.34	0.27	0.34
15.5	0.34	0.29	Not applicable
16.6	0.34	0.29	Not applicable

3.3.1 Taking into Account Nominal GDP and Population

Other option is to take into account nominal GDP and some regressors related to the steady-state, such as the Spanish population growth rate, since the estimation period falls within the long term. The following equation describes this case.

$$ENERGY\_DEMAND_t = \beta_0 + \beta_1 HDD_t + \beta_2 CDD_t + \beta_3 PREC_t + \beta_4 nominalGDP_t + \beta_5 log(\frac{SpanishPopulation_t}{SpanishPopulation_{t-1}}) \quad (14)$$

However, this model also does not improve the initial results as shown below (Tables 9 and 10).

Table 9 MSE adding information about nominal GDP and population growth rate

Base temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,091,721	166,109,399	940,805
15.5	1,088,966	7,652,457	1,036,323
16.6	1,086,226	432,212,828	1,289,250

Table 10 R^2 adding information about nominal GDP and population growth rate

Base temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.37	Not applicable	0.46
15.5	0.37	Not applicable	0.40
16.6	0.37	Not applicable	0.26

Interpreting the results and using the linear model and the Neural Networks, we can conclude that coefficients for HDD, CDD, and intercept parameter are significants (p-value less than 0.001). SVR obtain the worse results with this model. We do not find additional relations.

4 Simulating Energy Demand for 2020

We simulated a scenario that assumes that the IPI in Spain will grow continuously in the future, repeating previous growth rates. The simulation process takes into account the following hypothetical scenario:

1. IPI: It is assumed that Spain began a process of economic activation resulting in an increase in industrial activity. Therefore from January 2016 we will begin repeating (backwards) the data we have from December 2011 to November 2008. This implies the assumption that, in December 2019, the value of IPI will be 127.39. Our assumption is based on the evolution of IPI over the last years. The average annual growth rate for the period January 2014 to November 2016 (last data available) has been 2.24. The average annual growth rates for 2015 and 2016 have been3.24 and 2.04, respectively. These values lead us to trust our assumption.
2. Climatological data: We have assumed that weather will not change over the previous two years. Therefore, for each month and for the HDD, CDD and rainfall variables we take means of the corresponding month in the last two years. CDD and HDD have been calculated with a base temperature value of 15.5 °C.

We are not assuming the direction of the time series relationship. We are just assuming a counterfactual and we try to evaluate our assumptions using this counterfactual.

Always according to the proposed scenario, the Deep Learning Model estimates that energy demand in December 2019 will be 23,109.17 GWh. If we use a base temperature of 14.5 °C the prediction for energy demand in December 2019 is 22,818.62 GWh while with a value of comfort of 16.5 for the base temperature, the expected energy demand according to the model's prediction will be 24,364.65 GWh.

The amount of CO_2 emitted is important for its environmental impact. Therefore, this paper also includes a small exercise about this issue. According to the Electricity Observatory data of World Wildlife Fund (WWF), in December 2015, Spain had issued an average of 0.269 kg of CO_2 per KWh consumed. Thus, if we consider an expected demand for December 2019 in the average scenario of 23,109.17 GWh, and assuming that the ratio of CO_2 emissions is maintained, Spain will emit into the atmosphere more than 6.2 million tons of CO_2. Introducing uncertainty in the scenarios based on the temperature of comfort, the interval of emissions will move from 6.1 to 6.5 million tons. Given that the Paris agreement on emissions will enter into force by the end of 2016, there is place for technology, for the energy mix and for efficiency to achieve our commitments.

5 Conclusions

We can observe the following facts in the demand for energy in Spain in recent years:

1. Since March 2007, according to available data, the monthly energy demand has been higher than the amount we simulated for December 2019 only 11 times. But since January 2013, only in July 2015 energy demand exceeded this level, reaching a value of 23,476 GWh.
2. According to a report by REE, in July 2015, generated of electricity from renewable energy sources reached 30.7% of the total energy produced. However, this figure includes renewable thermal energy (1.8%), which is obtained by burning waste and is thus a contaminant. In any case, there is a potential to reach the goal of 30% energy produced from renewable sources.

In 2005 electricity demand in the Iberian Peninsula amounted to 246,187 GWh. While electrical consumption in the islands was 14,517 GWh amounting to a total of 260,704 GWh for the whole of Spain. On the other hand, we believe that the ratio of CO_2 emissions per KWh consumed in 2005 is higher than the ratio expected for 2019. This is because electricity technologies continue evolving towards more sustainable production methods, for example, increasing renewable energy sources. If the emissions in 2019 is 0.269 kg of CO_2 per KWh consumed) we predict that the volume of CO_2 emissions at the end of 2019 will correspond to around 75 million tons (between 6.1 and 6.5 million tons per month).

We think that this kind of exercise evidence illustrates the need for detailed, downloadable, easily usable and validated open data about energy consumption and emission that allows us to analyze whether the initial objectives about greenhouse gas emissions are going to be achieved. As a preliminary conclusion, our evidence suggests that Spain may be on track to meet its commitments to European Union. These agreements imply that, by 2020 between 20 and 30% of the consumed energy comes from clean and renewable energy sources. We have serious doubts about how much could be achieved in reducing pollutant gases. However, the clean and renewable energy source development can be achieved by external factors to the regulation itself or, by developing energy policies aimed at the introduction of a sustainable mix (following the path initiated in the early 2000s), by the development of real sectors energy-related where Spain was considered highly innovative in terms of technology and by consumer awareness with improving energy efficiency. This productive model should boost activity and employment in relation to energy efficiency and clean energy production.

Acknowledgements We appreciate the comments received during the presentation of this work at the 4th International Conference on Dynamics, Games and Science: Decision models in a complex economy held at Universidad Nacional de Educación a Distancia (UNED), Madrid. We would like to thank all people involved in organizing this conference. In particular, we would like to thank the Organizing Committee members, colleagues, editors and referees of the paper.

References

1. Capros, P., Mantzos, L., Parousos, L., Tasios, N., Klaassen, G., Van Ierland, T.: Analysis of the EU policy package on climate change and renewables. Energy Pol. **39**(3), 1476–1485 (2011)
2. Red Eléctrica de España. Spain Electrical Energy Report (in Spanish). May 2016. http://www.ree.es/sites/default/files/downloadable/inf_sis_elec_ree_2015.pdf
3. López, C.B., Carreras, A., Tafunell, X.: Estadísticas históricas de Espaa: siglos XIX-XX. **(3)**. Fundación BBVA. Madrid (2005)
4. Tiana, M.: El impacto de la crisis económica sobre la industria espaola. Boletín Económico. Banco de Espaa, Madrid (2012)
5. National Centers for Environmental Information. National Oceanic and Atmospheric Administration (1999). http://www.ncdc.noaa.gov/. Cited 18 Jun 2016
6. Ristinen, R.A., Kraushaar, J.J.: Energy and the Environment. Wiley-VCH, New York (1998)
7. Labandeira, X., Labeaga, J.M., Lpez-Otero, X.: A meta-analysis on the price elasticity of energy demand. Energy Pol. **102**, 549–568 (2017)
8. Wolpert, D.H., Macready, W.G.: No free lunch theorems for search. **10** Technical Report SFI-TR-95-02-010, Santa Fe Institute (1995)
9. Team, R.C.: R: A language and environment for statistical computing (2013)
10. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 6. Springer, New York (2013)
11. Stigler, S.M.: Gauss and the invention of least squares. Ann. Stat. **9**, 465–474 (1981)
12. Hayashi, F.: Econometrics. Princeton University Press, Princeton (2000)
13. Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W.: Applied Linear Statistical Models, vol. 4. Irwin, Chicago (1996)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. learn. **20**(3), 273–297 (1995)
15. Smola, A.J., Schlkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)
16. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. **26**, 1651–1686 (1998)
17. Chapelle, O., Vapnik, V.: Model selection for support vector machines. In: NIPS, pp. 230–236 (1999)
18. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071) (2015). https://CRAN.R-project.org/package=e1071
19. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
20. Arel, I., Rose, D.C., Karnowski, T.P.: Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE Comput. Intell. Mag. **5**(4), 13–18 (2010)
21. Zhang, G., Patuwo, B.E., Hu, M.Y.: Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. **14**(1), 35–62 (1998)
22. Deaton, A., Muellbauer, J.: Economics and Consumer Behavior. Cambridge University Press, Cambridge (1980)

Simulation and Advanced Control of the Continuous Biodiesel Production Process

Ana S. R. Brásio, Andrey Romanenko and Natércia C. P. Fernandes

Abstract The biodiesel industry is characterized by high fluctuations of the prices and a multiplicity of biological raw material sources. On the other hand, there exist strict quality standards imposed on the final product. Because of these factors, it is important for biodiesel plants to run their processes in the most efficient manner in order to stay competitive. One of the ways to achieve this is the use of model based approaches for design, operation, and control. In this work, that focuses on the latter two areas, a first-principle dynamic model of the main units of a biodiesel plant is developed and applied in two situations: for open-loop simulation as well as for process optimization. The former demonstrates the response observed in the process variables when the plant is subjected to a series of disturbances in the input variables. The later is built in the context of nonlinear model predictive control that determines the optimal profiles of the manipulated variables taking into account process and quality constraints as well as the associated reactant and energy costs.

Keywords Continuous biodiesel production · Process modeling · Nonlinear model predictive control

Nomenclature

c_p^*	molar specific heat capacity	J mol^{-1} °C^{-1}
C	molar concentration	mol m^{-3}
d	vector of disturbances	various
E_a	activation energy	J mol^{-1}
F	mass flow rate	kg s^{-1}
n	molar amount of molecules	mol
n_d	number of disturbance variables	dimensionless

A. S. R. Brásio
CIEPQPF, Department of Chemical Engineering, University of Coimbra, Portugal & Ciengis, Coimbra, Portugal

A. Romanenko
Ciengis, Coimbra, Portugal

N. C. P. Fernandes (✉)
CIEPQPF, Department of Chemical Engineering, University of Coimbra, Portugal, Coimbra, Portugal
e-mail: natercia@eq.uc.pt

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_6

n_m	number of control variables	dimensionless
n_o	number of output variables	dimensionless
n_s	number of state variables	dimensionless
n_θ	number of parameter variables	dimensionless
k	specific reaction rates constants	$m^3\ mol^{-1}\ s^{-1}$
k_0	pre-exponential factor	$m^3\ mol^{-1}\ s^{-1}$
m	control horizon length	dimensionless
M	molar mass	$kg\ mol^{-1}$
N	molar flow rate	$mol\ s^{-1}$
N'	molar flow rate of the final biodiesel stream	$mol\ s^{-1}$
p	predictive horizon length	dimensionless
r	overall reactions rates	$mol\ m^3\ s^{-1}$
R	the ideal gas constant	$J\ mol^{-1}\ °C^{-1}$
V	total volume	m^3
t	continuous time	s
T	temperature	°C
u	vector of manipulated variables	various
u^*	vector of optimal manipulated variables	various
U	augmented vector of initial control profiles	various
U^*	augmented vector of optimal initial control profiles	various
x	molar fraction	$mol\ mol^{-1}$ *
x'	molar fraction of the final biodiesel stream	$mol\ mol^{-1}$
w	weighting scalars	various
y	vector of output variables	various
y^*	vector of optimal output variables	various
y'	mass fraction of the final biodiesel stream	$kg\ kg^{-1}$ *
Y	augmented vector of initial output profiles	various
$\tilde{Y}$	augmented vector of output predictions	various
z	vector of state variables	various
z^*	vector of optimal state variables	various
Z	augmented vector of initial state profiles	various
Z^*	augmented vector of optimal initial state profiles	various
$\tilde{Z}$	augmented vector of state predictions	various
$\tilde{Z}^*$	augmented vector of optimal state predictions	various

* When explicitly stated, the values of y'_E and of $x_{lt,M}$ might be expressed in $\%(m/m)$ and $\%(n/n)$, respectively

ΔH_r	heat of reaction	$J\ mol^{-1}$
Δt	sampling time	s
$\mathcal{H}$	Heaviside function	dimensionless
ρ	density	$kg\ m^{-3}$
Ψ	cost function	dimensionless
ϕ	generation-reaction term	$mol\ m^{-3}\ s^{-1}$
θ	vector of parameters	various
ξ	split fraction of the input component to the light phase	dimensionless

Acronyms:

TG triglycerides
DG diglycerides
MG monoglycerides
E esters of fatty acids

G glycerol
M methanol
O oil

Subscripts:

H heat exchanger
R reactor
D decanter
hv heavy phase in the decanter
lt light phase in the decanter
L lower bound
U upper bound
sp setpoint
ref reference

1 Introduction

The operational costs (which include raw materials, utilities, labour, supplies, and general work) constitute a major part of overall biodiesel production costs [6]. This fact is perhaps in the basis of the high fluctuations of biodiesel prices and in the profitability margins.

In spite of the diversity and the high variability of the raw material that ranges from virgin or waste vegetable oils to algae and even to animal fats [12], commercial biodiesel has to comply with standard requirements in what concerns its composition and characteristics. In particular, the standard EN 14 214 [7] specifies that a minimum of 96.5% (m/m) of fatty acid methyl esters should be present in the final product.

In this context, the economic success of any biodiesel plant depends on its ability to operate the production processes in the most efficient way. Although model based solutions for operation and control of biodiesel plants are of importance in this regard, the use of such tools in biodiesel industry is not widespread yet. Some contributions in this area may be found in the literature [1, 3–5, 11, 15].

The present work provides a holistic model based approach for the operation of a continuous biodiesel production line. A model of the line (constituted by the key units reactor, heat exchanger, decanter, washer, and dryer) is developed taking into account the specifics of the control and optimization context. A control strategy is designed to ensure an efficient operation while complying with the stringent quality

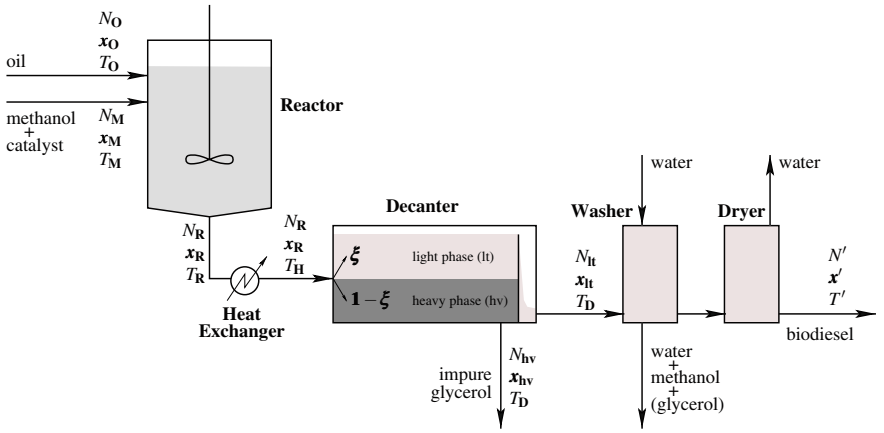


Fig. 1 Schematic representation of the system

requirements. A use case demonstration is developed to show the benefits of the approach.

2 Mathematical Model

2.1 The System

The system depicted in Fig. 1 transforms a biological raw material (oil) into biodiesel. It consists on five operational units installed in series: a reactor, a heat exchanger, a decanter, a washer, and a dryer, working in continuous mode.

Oil and methanol enter the reactor, where appropriate mixing ensures an effective contact between oil triglycerides (TG) and methanol (M). Under convenient temperature conditions and the action of a catalyst, the transesterification reaction between TG and M occurs yielding esters of fatty acids (E) and the sub-product glycerol (G). The overall reaction comprises three steps [13]:

$$\mathrm{TG} + \mathrm{M} \underset{\mathrm{k_2}}{\overset{\mathrm{k_1}}{\rightleftharpoons}} \mathrm{DG} + \mathrm{E}$$

$$\mathrm{DG} + \mathrm{M} \underset{\mathrm{k_4}}{\overset{\mathrm{k_3}}{\rightleftharpoons}} \mathrm{MG} + \mathrm{E}$$

$$\mathrm{MG} + \mathrm{M} \underset{\mathrm{k_6}}{\overset{\mathrm{k_5}}{\rightleftharpoons}} \mathrm{G} + \mathrm{E}\,,$$

where DG and MG stand for di and monoglyceride, respectively. The reacting mixture exits continuously the reactor at a flow rate that warranties a constant level in the unit. Since a methanol to oil molar ratio around 6:1 is typically used (i.e., there is excess of methanol), the mixture leaving the reactor contains mainly ester, glycerol, and methanol together with residual amounts of not reacted tri, di, and monoglycerides.

The required separation between ester and glycerol is difficult at the relatively high temperature of the mixture that exits the reactor. Therefore, that stream is cooled down in a heat exchanger before it enters the decanter, where the separation takes place.

In the absence of mixing and under the action of gravity, the immiscible compounds E and G tend to go apart forming two individualized phases: a light (less dense) upper phase containing almost all the ester and a heavy (more dense) down phase containing almost all the glycerol. Methanol splits over both phases while virtually all residual glycerides migrate to the light phase because of their affinity with ester molecules. The down heavy phase (hv) leaves the bottom of the decanter at a flow rate that ensures a constant level of the heavy phase inside the unit. The upper light phase (lt) exits the decanter by overflowing a baffle located near its end.

The crude biodiesel exiting the decanter as the light phase is then sent to the washer to remove any remaining methanol, glycerol, or catalyst. Finally, the crude biodiesel undergoes the process of drying that removes the remaining water to the level that is compliant with the biodiesel quality specification.

2.2 The Model

Reactor

A mathematical model that describes the behavior of the reactor can be obtained by applying mass and energy balances to the reaction unit. Such balances take into consideration the chemical and physical phenomena happening in the reactor (pointed out in Sect. 2.1).

With the purpose of writing the model equations in a more compact way, two sets are defined as

$$\mathrm{S} = \{\text{oil, methanol}\} = \{\mathrm{O}, \mathrm{M}\},$$
$$\mathrm{I} = \{\mathrm{M}, \mathrm{TG}, \mathrm{DG}, \mathrm{MG}, \mathrm{G}, \mathrm{E}\},$$

which are the set of input streams in the reactor and the set of chemical species involved, respectively.

Assuming that the mixture is perfect, the evolution of the liquid inside the reactor in terms of its composition (expressed by the molar fractions of the six chemical species) and temperature can be predicted by

$$n_{\mathrm{R}}\,\frac{\mathrm{d}x_{\mathrm{R},i}}{\mathrm{d}t} = \sum_{s\in\mathrm{S}} N_s\,(x_{s,i} - x_{\mathrm{R},i}) + \phi_i \qquad (i \in \mathrm{I})\,, \tag{1a}$$

$$n_{\mathrm{R}}\,c^{*}_{\mathrm{p,R}}\,\frac{\mathrm{d}T_{\mathrm{R}}}{\mathrm{d}t} = \sum_{s\in\mathrm{S}} N_s\,c^{*}_{\mathrm{p},s}\,(T_s - T_{\mathrm{R}}) + \sum_{j=1}^{3}\big(-\Delta H_{\mathrm{r}}\big)_j\,r_j\,V_{\mathrm{R}}\,, \tag{1b}$$

with the total molar amount of molecules in the mixture that is inside the reactor, n_{R}, given by

$$n_{\mathrm{R}} = V_{\mathrm{R}}\cdot\Big(\sum_{i\in\mathrm{I}}\frac{\mathrm{M}_i}{\rho_i}\,x_{\mathrm{R},i}\Big)^{-1},$$

and the generation-reaction terms, ϕ_i, defined as

$$\begin{aligned}
&\phi_{\mathrm{M}} = -(r_1 + r_2 + r_3)V_{\mathrm{R}}\,, \quad && \phi_{\mathrm{TG}} = -r_1 V_{\mathrm{R}}\,,\\
&\phi_{\mathrm{DG}} = (r_1 - r_2)V_{\mathrm{R}}\,, && \phi_{\mathrm{MG}} = (r_2 - r_3)V_{\mathrm{R}}\,,\\
&\phi_{\mathrm{G}} = r_3 V_{\mathrm{R}}\,, && \phi_{\mathrm{E}} = (r_1 + r_2 + r_3)V_{\mathrm{R}}\,.
\end{aligned}$$

The overall reactions rates associated to the three reaction steps, r_j, are given by

$$r_1 = k_1 C_{\mathrm{TG}} C_{\mathrm{M}} - k_2 C_{\mathrm{DG}} C_{\mathrm{E}}\,,\quad r_2 = k_3 C_{\mathrm{DG}} C_{\mathrm{M}} - k_4 C_{\mathrm{MG}} C_{\mathrm{E}}\,,\quad r_3 = k_5 C_{\mathrm{MG}} C_{\mathrm{M}} - k_6 C_{\mathrm{G}} C_{\mathrm{E}}\,,$$

where k_l represents the specific reaction rate constant of reaction l and C_i the molar concentration of component i in the reactor mixture. The specific reactions rates constants are calculated from the Arrhenius equation

$$k_l = k_{0l}\,\exp(-E_{\mathrm{a},l}/(R\,T_{\mathrm{R}})) \qquad (l = \{l \in \mathbb{N} : 1 \leqslant l \leqslant 6\})\,,$$

where $k_{0,l}$ is the pre-exponential factor, $E_{\mathrm{a},l}$ is the activation energy, R is the ideal gas constant, and T_{R} is the temperature of the mixture inside the reactor. The molar concentration of chemical species i in the reactor can be expressed in terms of its counterpart molar fraction according to

$$C_i = x_{\mathrm{R},i}\cdot\Big(\sum_{i\in\mathrm{I}}\frac{\mathrm{M}_i}{\rho_i}\,x_{\mathrm{R},i}\Big)^{-1} = \frac{n_{\mathrm{R}}}{V_{\mathrm{R}}}\,x_{\mathrm{R},i}\,.$$

From a global mass balance, and in order to ensure the assumption of constant height in the reactor, the molar flow leaving the reactor, N_{R}, is

$$N_{\mathrm{R}} = \frac{1}{M_{\mathrm{R}}}\sum_{s\in\mathrm{S}} M_s N_s - \frac{V_{\mathrm{R}}\sum_{i\in\mathrm{I}} M_i\frac{\mathrm{d}x_{\mathrm{R},i}}{\mathrm{d}t}}{M_{\mathrm{R}}\sum_{i\in\mathrm{I}}\frac{\mathrm{M}_i}{\rho_i}x_{\mathrm{R},i}} + \frac{V_{\mathrm{R}}\sum_{i\in\mathrm{I}}\frac{\mathrm{M}_i}{\rho_i}\frac{\mathrm{d}x_{\mathrm{R},i}}{\mathrm{d}t}}{(\sum_{i\in\mathrm{I}}\frac{\mathrm{M}_i}{\rho_i}x_{\mathrm{R},i})^2}\,,$$

Table 1 Thermo-physical properties of each component at 60 °C [5]

	Units	M	TG	DG	MG	E	G
ρ	kg/m^3	757	954	983	1030	844	1340
c_p^*	J/(mol °C)	2785	2110	2188	2381	2146	2556
M	10^{-3} kg/mol	32	853	600	346	286	92

Table 2 Heat of reaction at 60 °C [5]

Units	$(\Delta H_r)_1$	$(\Delta H_r)_2$	$(\Delta H_r)_3$
J/mol	15699	36899	−58906

with the average molar mass of a liquid mixture given by the weighted average of the molar masses of its individual components, M_i ($i \in$ I), that is,

$$M = \sum_{i \in \mathrm{I}} x_i \, M_i \ .$$

The molar specific heat capacity of a liquid mixture can be obtained directly from the specific heat capacity of the individual components of that mixture weighted by their corresponding molar fractions. Therefore, for a generic mixture

$$c_p^* = \sum_{i \in \mathrm{I}} x_i \, c_{p,i}^* .$$

In what concerns TG, DG, MG, and E, a pseudo-component approach explained somewhere else [3] was adopted to take into consideration the diversity of chemical compounds present in the oil feed due its biological nature. Assuming an oil composition identical to that used in [5], the molar specific heat capacity, c_p^*, can be directly obtained from the specific heat capacity and the molar mass also indicated in [5]. The values of these physical properties for each chemical species at a temperature of 60 °C are indicated in Table 1.

Also, for the same oil composition, the heats of reaction were computed in [5]. These values are shown in Table 2.

Although these properties are somewhat dependent on temperature, such effect is not perceptible in the overall system. Thus, all the mentioned parameters (ρ, c_p^*, and (ΔH_r)) were considered constant for the operating conditions range used in this work.

Finally, the activation energies and the pre-exponential factors were estimated based on the kinetic experimental studies of [9] for the palm oil. The Arrhenius equation was adjusted to the experimental reaction rates constants obtained for temperatures 40, 50 and 60 °C. Two experimental points were eliminated for lacking physical meaning. According to the experimental data, the last reverse reaction rate

constant (k_6) for palm oil appears to be independent of temperature and the reaction itself is practically inexistent, that is, the ester does not get converted into monoglyceride. Thus, both parameters $k_{0,6}$ and E_a were made zero. Table 3 contains the values of the adjusted kinetic parameters.

Heat Exchanger

By withholding an appropriate amount of energy, the heat exchanger decreases the temperature of the stream that leaves the reactor to the desired value. Simultaneously, since there is no mass transfer inside the heat exchanger, the mass fractions of the stream that leaves this unit are exactly the same as those of the input stream (i.e., the stream that exits the reactor).

Decanter

A previous work on the modelling of a continuous decanter was presented in [5]. Since that work studied exclusively the decanter unit, it took into consideration the three main chemical species from the point of view of this unit: E, G, and M. However, the broader system under study in the present work requires to take other components into consideration since they have a rather important role in the first unit of the system (the reactor): TG, DG, and MG. Therefore, the first principle model presented in [4] was extended from three to six chemical species. There was also need to incorporate the temperature of the decanter as a variable, since the temperature of its feed is dependent on the reactor unit that is now part of the considered system. The extended dynamic model of the decanter can be written (for $i \in \mathrm{I}\backslash\{\mathrm{G}\}$) as

$$n_{\mathrm{hv}} \frac{\mathrm{d}x_{\mathrm{hv},i}}{\mathrm{d}t} = \sum_{k\in\mathrm{I}} \left((1-\xi_k)\, x_{\mathrm{R},k}\right) N_{\mathrm{R}} \left(\frac{1-\xi_i}{\sum_{k\in\mathrm{I}} \left((1-\xi_k)\, x_{\mathrm{R},k}\right)} x_{\mathrm{R},i} - x_{\mathrm{hv},i} \right) \tag{2a}$$

$$n_{\mathrm{lt}} \frac{\mathrm{d}x_{\mathrm{lt},i}}{\mathrm{d}t} = \sum_{k\in\mathrm{I}} \left(\xi_k\, x_{\mathrm{R},k}\right) N_{\mathrm{R}} \left(\frac{\xi_i}{\sum_{k\in\mathrm{I}} (\xi_k\, x_{\mathrm{R},k})} x_{\mathrm{R},i} - x_{\mathrm{lt},i} \right) , \tag{2b}$$

where n_{hv} and n_{lt} represent the molar amount of molecules in the heavy and light phases, respectively. The composition of phase j (with $j = \{\mathrm{hv}, \mathrm{lt}\}$) in the remaining component (G) is

$$x_{j,\mathrm{G}} = 1 - \sum_{i\in\mathrm{I}\backslash\{\mathrm{G}\}} x_{j,i} \ . \tag{3}$$

A mathematically equivalent alternative to (3) consists on applying (2) also for the component glycerol.

The split fraction, ξ_i, quantifies the division of component i by the two phases. It represents the fraction of the input component i that goes to the light phase and complementary $1 - \xi_i$ indicates the fraction that goes to the heavy phase. The split fractions depend on the instantaneous composition and temperature of the stream that is continuously entering the decanter. The mechanistic models to quantify the

Table 3 Kinetic parameters based on experimental data of [9] for palm oil

Parameter	Units	k_1	k_2	k_3	k_4	k_5	k_6
E_a	J/mol	30225	44646	28941	72779	56093	0
k_0	L/(mol min)	1.3613×10^{-1}	4.6645×10^{1}	3.8708×10^{-1}	1.4710×10^{7}	4.5477×10^{4}	0

liquid-liquid equilibrium are iterative and thus not appropriate in the context of the present work, as discussed in [5]. A neural network model that is able to substitute and avoid the iterative algorithm in calculating the split fractions for E, G, and M was developed previously [5] for compounds E, G, and M and is adopted here in order to compute ξ_{M}, ξ_{G}, and ξ_{E}. In what concerns the compounds TG, DG, and MG, they are present only in residual amounts. Moreover, their strong affinity with ester molecules makes virtually all of the glycerides migrate to the light phase. Based on these circumstances, the split fractions for TG, DG, and MG were considered to be equal to 1.

To ensure constant height of the heavy phase (and, therefore, constant height also of the light phase since the total height of the overflow baffle is constant), the variation of the molar amount of molecules in phase j (with $j = \{\mathrm{hv}, \mathrm{lt}\}$) is

$$\frac{\mathrm{d}n_j}{\mathrm{d}t} = -\frac{1}{\sum\limits_{i\in\mathrm{I}}\left(\frac{\mathrm{M}_i}{\rho_i}\, x_{j,i}\right)}\, n_j \sum_{i\in\mathrm{I}}\left(\frac{\mathrm{M}_i}{\rho_i}\frac{\mathrm{d}x_{j,i}}{\mathrm{d}t}\right) .$$

From an energy balance to the decanter unit, it is possible to write

$$(n_{\mathrm{lt}} + n_{\mathrm{hv}})\, c^*_{\mathrm{p,D}}\, \frac{\mathrm{d}T_{\mathrm{D}}}{\mathrm{d}t} = N_{\mathrm{R}}\, c^*_{\mathrm{p,R}}\, (T_{\mathrm{H}} - T_{\mathrm{D}}) , \tag{4}$$

neglecting the heat exchanges between the liquid inside the decanter and the environment, since the operating temperature of the decanter is not very different from room temperature.

Washer + Dryer

In higher capacity plants, crude biodiesel is typically purified using water. In fact, the washer unit removes residual methanol and glycerol as well as remaining sodium salts and soaps. Because of their affinity with water molecules, these impurities get retained in the washing liquid and the resulting stream gets easily separated from biodiesel due to the difference in the density. Several washing cycles might be needed. After the washing process, the crude biodiesel is dried in order to bring the remaining water and methanol concentration in accordance with the quality specification. Assuming that the washing and the drying are perfect, all the methanol and residual glycerol that exists in the upper stream that leaves the decanter are retained by the washing water while all the other components (E, and residual TG, DG, and MG) remain untouched and, also, all the remaining water is dried up. Therefore, the composition of the mixture after washing and drying (which constitutes the biodiesel) is given by the molar mass fractions

$$x'_i = \frac{x_{\mathrm{lt},i}}{x_{\mathrm{lt,TG}} + x_{\mathrm{lt,DG}} + x_{\mathrm{lt,MG}} + x_{\mathrm{lt,E}}} \qquad (i \in \mathrm{I}\backslash\{\mathrm{M}, \mathrm{G}\}) . \tag{5}$$

The composition of the final biodiesel stream can be expressed in mass fraction by

$$y'_i = \frac{x'_i\, M_i}{\sum\limits_{j \in \mathrm{I}\backslash\{\mathrm{M,G}\}} x'_j\, M_j} = \frac{x_{\mathrm{lt},i}\, M_i}{\sum\limits_{j \in \mathrm{I}\backslash\{\mathrm{M,G}\}} x_{\mathrm{lt},j}\, M_j} \qquad (i \in \mathrm{I}\backslash\{\mathrm{M,G}\})\ . \tag{6}$$

All the mass or molar fractions can be equivalently expressed as mass or molar percentage, respectively, by simply multiplying the former by 100%.

Finally, a generic molar flow rate N can be expressed as a mass flow rate F using the average molar mass of the stream it refers to, that is,

$$F = M\, N\ .$$

3 Control Problem Formulation

In order to produce biodiesel that complies with the standard quality requirement of 96.5% (mass percentage) in ester, the continuous system described above is subjected to the action of an advanced control scheme in a holistic approach. Moreover, the costs associated with the amount of water used in the purification of the raw biodiesel and the subsequent methanol recovery make it preferable to redirect the non-reacted methanol to the down phase in detriment of the light upper phase. Thus, the molar fraction of M present in the decanter light phase is also controlled. The control of the variables y'_{E} and $x_{\mathrm{lt,M}}$ is achieved by manipulating the inlet flow rate of reactant methanol, F_{M}, the temperature of the oil feed, T_{O}, and the output temperature of the heat exchanger, T_{H}. The objective of the nonlinear model predictive control (NMPC) used is to determine, for a certain predictive horizon, the optimal profiles of the manipulation variables that allow to satisfy the setpoints of the controlled variables and the operating constraints.

The control problem is configured with 2 output variables (vector y) and 3 manipulated variables (vector u) as listed in Table 4. The NMPC implemented relies on the

Table 4 NMPC configuration parameters

Variable	Units	LB	UB	RB	Setpoint	Reference	Weight
		L	U	Δu_k	$y_{\mathrm{sp},k}$	$u_{\mathrm{ref},k}$	w
Controlled variables (y)							
y'_{E}	$\%(m/m)$	96.5	100.0	–	97.6	–	10^4 (kg/kg)$^{-2}$ 10^{-4}
$x_{\mathrm{lt,M}}$	$\%(n/n)$	0.0	30.0	–	24.4	–	10^1 (mol/mol)$^{-2}10^{-4}$
Manipulated variables (u)							
F_{M}	kg/h	0	657	±100	–	657	10^{-2} h^2 kg^{-2}
T_{O}	°C	25	63	±5	–	Floating	10^{-2} °C^{-2}
T_{H}	°C	25	60	±5	–	Floating	10^{-1} °C^{-2}

LB – lower bound
UB – upper bound
RB – rate bound

process model described by Eqs. (1)–(5) and on the observation model given by (6) which can be compactly represented by

$$\dot{z} = f(z, u, d, \theta)\,, \tag{7a}$$
$$y = g(z)\,, \tag{7b}$$

with f and g twice continuously differentiable, where $z \in \mathbb{R}^{n_s}$ is the vector of state variables, $u \in \mathbb{R}^{n_m}$ is the control vector, $d \in \mathbb{R}^{n_d}$ is the disturbance vector, $\theta \in \mathbb{R}^{n_\theta}$ is the parameter vector, and $y \in \mathbb{R}^{n_o}$ is the vector of output variables.

The continuous time t is discretized via a sampling time Δt and a time instant index k ($k \in \mathbb{N}$). At every time index k, the following NMPC problem is solved: knowing the process and the observation models as well as the current state measurements and/or estimations, z_k, compute, based on a predictive horizon of length p, the optimal control input sequence over a control horizon of length m for $m \leqslant p$, denoted by $\{u_k^*, \ldots, u_{k+m-1}^*\}$, with $u_{k+i-1}^* = u_{k+m-1}^*$ for $m < i \leqslant p$. Such optimal control input will lead to a sequence of state and output predictions over p that minimize the objective cost function of the NMPC problem, the optimal sequences designated by $\{z_{k+1}^*, \ldots, z_{k+p}^*\}$ and $\{y_{k+1}^*, \ldots, y_{k+p}^*\}$, respectively.

The NMPC problem is formulated as the discrete-time constrained dynamic optimization problem

$$\min_{X_k, U_k} \ \Psi\left(\tilde{Y}_k, U_k\right) \tag{8a}$$
$$\text{s.t. } \tilde{z}_{k+i} = \mathrm{f}(z_{k+i-1}, u_{k+i-1}, d, \theta, \Delta t) \qquad (1 \leqslant i \leqslant p) \tag{8b}$$
$$\tilde{y}_{k+i} = \mathrm{g}(\tilde{z}_{k+i}) \qquad (1 \leqslant i \leqslant p) \tag{8c}$$
$$u_{k+i-1} = u_{k+m-1} \qquad (m < i \leqslant p) \tag{8d}$$
$$z_{k+i} - \tilde{z}_{k+i} = 0 \qquad (1 \leqslant i < p) \tag{8e}$$
$$\tilde{Z}_\mathrm{L} \leqslant \tilde{Z}_k \leqslant \tilde{Z}_\mathrm{U}\,, \quad \tilde{Y}_\mathrm{L} \leqslant \tilde{Y}_k \leqslant \tilde{Y}_\mathrm{U} \tag{8f}$$
$$Z_\mathrm{L} \leqslant Z_k \leqslant Z_\mathrm{U}\,, \quad U_\mathrm{L} \leqslant U_k \leqslant U_\mathrm{U} \tag{8g}$$

where the augmented vectors of the output predictions, of the state predictions, of the initial state profiles, and of the initial control profiles are defined by $\tilde{Y}_k^\mathrm{T} = \left[\tilde{y}_{k+1}^\mathrm{T}, \cdots, \tilde{y}_{k+p}^\mathrm{T}\right]$, $\tilde{Z}_k^\mathrm{T} = \left[\tilde{z}_{k+1}^\mathrm{T}, \cdots, \tilde{z}_{k+p}^\mathrm{T}\right]$, $Z_k^\mathrm{T} = \left[z_k^\mathrm{T}, \cdots, z_{k+p-1}^\mathrm{T}\right]$, and $U_k^\mathrm{T} = \left[u_k^\mathrm{T}, \cdots, u_{k+m-1}^\mathrm{T}\right]$, respectively.

In this formulation, a multiple shooting strategy similar to that in [14] is used to perform the state predictions. This requires the set of equality constraints (8e) in order to guarantee the continuity of the state variables profiles over the predictive horizon. In (8e), $\tilde{z}_{k+i}$ is the state vector at $k+i$ obtained through the integration of the dynamic model inside each sampling time interval, with $t \in [t_{k+i-1}, t_{k+i}]$, using as initial conditions the nominal states and controls, z_{k+i-1} and u_{k+i-1}, respectively. Therefore, the decision variables of problem (8) are both the state and control trajectories, Z_k

and U_k, respectively. The subscripts $_{\mathrm{L}}$ and $_{\mathrm{U}}$ in the nonlinear constraints (8f) and in the decision variables bounds (8g) stand for *lower* and *upper* limit value, respectively.

This formulation can be interpreted as: at every time index k (correspondent to the sampling time $k\,\Delta t$) determine the optimal solution, Z_k^* and U_k^*, that minimizes the cost function (8a), such that the Z_k^* and $\tilde{Z}_k^*$ profiles match and are continuous over the predictive horizon, while satisfying all the problem constraints. The cost function is defined to be quadratic and involve the list of output variables to control and of input variables to manipulate given in Table 4 according to

$$\Psi\left(\tilde{Y}_k, U_k\right) = \sum_{\ell=1}^{2} w_\ell \sum_{i=1}^{p} \left(y_{\mathrm{sp},\ell,k+i} - \tilde{y}_{\ell,k+i}\right)^2 + \\ + \sum_{\ell=1}^{3} w_{2+\ell} \sum_{i=1}^{m} \left(u_{\mathrm{ref},\ell,k+i-1} - u_{\ell,k+i-1}\right)^2 , \tag{9}$$

where w_ℓ ($\ell \in \mathbb{N} : 1 \leqslant \ell \leqslant 5$) are weighting scalars indicated in Table 4. Other parameters of the control system configuration, such as the lower and upper bounds for controlled and manipulated variables, the rate bounds of manipulated variables (that is, the maximum change allowed for manipulated variables at each instant), the setpoint values for the controlled variables as well as the reference values considered for the manipulated variables, are listed in Table 4. It should be noted that other critical biodiesel quality and production cost parameters may be taken into account by the model predictive controller if the process model is extended to predict them.

In order to provide, at every time instant, the estimates of the state variables and some of the model parameters, it is employed the unscented Kalman filter (UKF) developed in [10]. The C++ based computational simulation framework Plantegrity® that was applied in this work features three independent and synchronized modules: the plant simulator module, the UKF module, and the NMPC module. The simulator and estimator modules use CVODES solver [8] for the solution of initial value problem. The nonlinear model predictive controller implements both multiple shooting [2] and simultaneous approaches for the solution of the underlying nonlinear constrained optimization problem using IPOPT [16]. A more detailed description of this framework is given in [3].

4 Results and Discussion

In consequence of its biological nature, the raw material of the process presents a rich diversity of components and, moreover, ordinary variations of its composition. Also, the final commercial product must comply with strict specifications legally imposed [7]. In order to ensure the minimum of 96.5% (m/m) required by the standard, the methanol could be used in larger excess pushing the reaction equilibria more to the right. However, the larger the excess of methanol used is, the bigger

are the costs associated to its further recovery. Still from a process point of view, the transesterification reaction is highly influenced by the reactor temperature. All these facts affect the system and therefore should be taken into consideration by the control mechanism to establish the way the system is run. The multivariable control configuration allows to produce biodiesel that is conforming to the product quality specification in what respects FAME content at the minimum specific cost related to energy and reactant (MeOH) consumption. It is also worth emphasizing that the kinetic model used in this work refers to the palm oil, which is a quite used raw material in biodiesel industry.

This discussion is organized in two parts. The first part reveals the behavior of the system in open-loop when stimulated by changes in its input variables. The second part exhibits the behavior of the system under the action of the developed control strategy.

4.1 Open-Loop Study

Figure 2 condenses the response of the system in four different scenarios, conveniently labeled from ① to ④, characterized by different stimuli:

① $F_{\mathrm{M}} = 657 - 331\,\mathcal{H}(t - 33.5) + 657\,\mathcal{H}(t - 67.0) - 326\,\mathcal{H}(t - 102.8)$;
② $T_{\mathrm{O}} = 60 - 3\,\mathcal{H}(t - 167.0) + 6\,\mathcal{H}(t - 202.7) - 3\,\mathcal{H}(t - 235.0)$;
③ $T_{\mathrm{H}} = 50 + 5\,\mathcal{H}(t - 301.3) - 10\,\mathcal{H}(t - 334.7) + 5\,\mathcal{H}(t - 370.5)$;
⑤ $F_{\mathrm{O}} = 3000 + 200\,\mathcal{H}(t - 434.7) - 400\,\mathcal{H}(t - 470.1) + 200\,\mathcal{H}(t - 500.0)$,

with $\mathcal{H}$ representing the Heaviside function. In all of them, the temporary stimulus consists of a pair of pulses (a negative followed by a positive or vice versa) so that the stimulated variable ultimately returns to its original value, as easily seen in the top graph of Fig. 2.

In response to the initial cut in the reactant methanol that occurs in scenario ①, the system evolutes with an abrupt decrease of the mass percentage of ester in the produced biodiesel, y'_{E}. In consequence, the final product rapidly falls in the off-specifications range (shadow area in the middle graph of Fig. 2). It is worth mentioning that in these circumstances not all the oil (TG, DG, and MG) is transformed into ester, as revealed by the top graph of Fig. 3. When the flow rate of methanol is later suddenly increased (the positive pulse of the disturbance), the relative amounts of glycerides and methanol in the reactor change significantly again, as the top graph of Fig. 3 evinces. The more prominent excess of methanol forces an increase of the extent of reactions leading to a practically complete transformation of the oil into ester, reason why the molar fractions of TG, DG, and MG in the reactor tend virtually to zero. Although the amounts of ester and of glycerol produced are now bigger, both the molar fractions of ester and of glycerol decrease because of the dilution effect originated by the bigger amount of methanol in the reactor. These changes are reflected ahead in the decanter, as shown by the middle and bottom graphs of Fig. 3. In particular, the ester content of the light phase that leaves the decanter suffers an

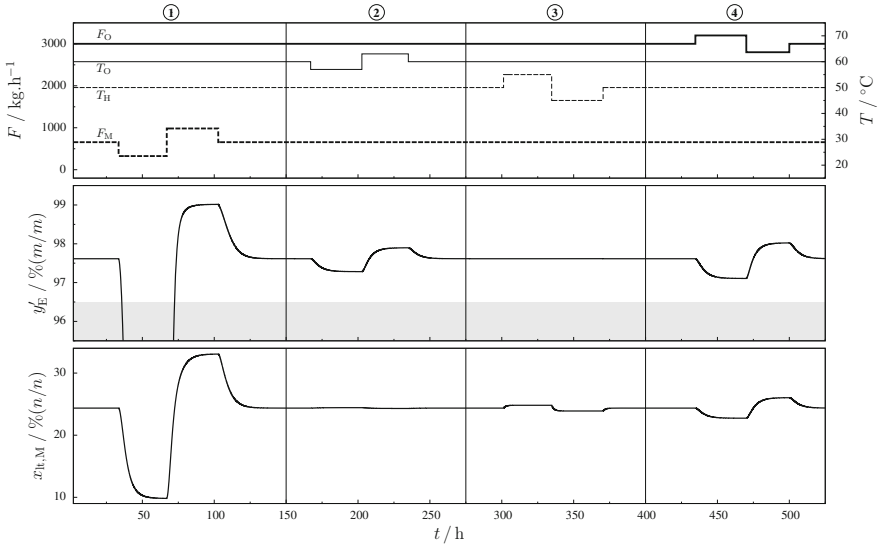


Fig. 2 Stimuli and response of key variables of the reactor-decanter system in open-loop

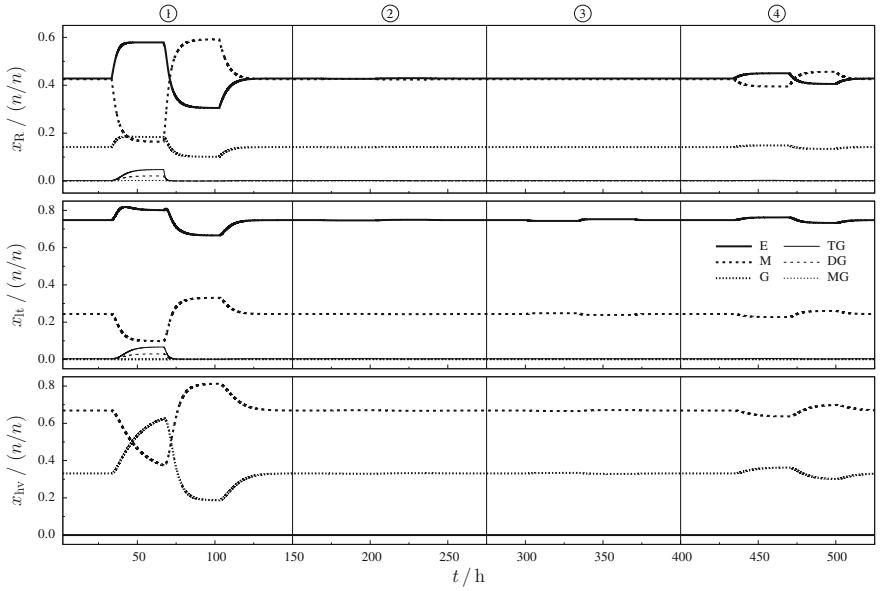


Fig. 3 Open-loop response of some of the state variables of the reactor-decanter system

evident decrease, since the non-reacted excess of methanol in the light phase disguises the real increase in the amount of ester in that stream. This disguising effect is eliminated after the withdrawal of methanol in the washer (see the bottom graph of Fig. 2). Thus, the mass percentage of ester in the final product increases (middle graph of Fig. 2). Both variables the mass percentage of ester in the final product and the molar percentage of methanol in the light phase are quite sensitive to changes in methanol flow rate. Once the load is finished, the system evolutes to the initial steady-state. It is noteworthy the nonlinear nature of the system, reflected by its asymmetric responses to symmetric stimuli. Although both negative and positive pulses in methanol flow rate have similar amplitudes, the amplitude of the corresponding responses is not the same, as it can be easily observed in the middle and bottom graphs of Figs. 2 and 4.

In scenarios ② and ③, the system is stimulated with changes in the temperature of the input oil stream and in the energy exchanged in the heat exchanger which ultimately corresponds to changes in the temperature of the stream that enters the decanter, respectively. The responses obtained in both scenarios ② and ③ are modest when compared to that of scenario ①. In spite of that, the stimuli were enough to originate considerable variations of y'_{E} in scenario ② and of $x_{\mathrm{lt,M}}$ in scenario ③.

Finally, in scenario ④, the effect of the oil flow rate on the system is studied. By analysing Figs. 2 and 3, it is possible to conclude that the system is rather sensitive to this input variable: increasing F_{O} pushes both y'_{E} and $x_{\mathrm{lt,M}}$ down. An increase in the input flow rate of oil affects the system in an opposite direction of an increase in the input flow rate of methanol. The effect is less pronounced than that verified in scenario ①, but the amplitude of the stimulus used in scenario ④ was also smaller.

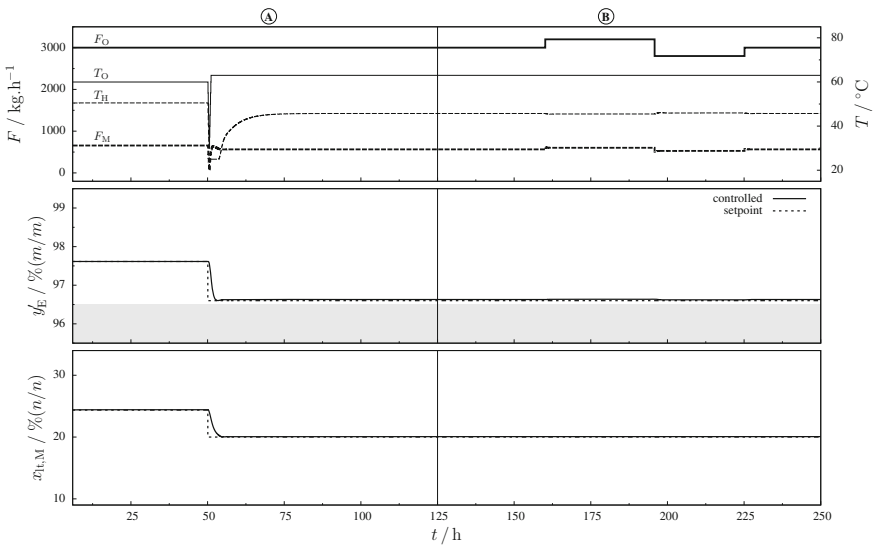


Fig. 4 Closed-loop response of the reactor-decanter system

4.2 Closed-Loop Study

The same system is now submitted to the action of the controller described in Sect. 3. In a first stage (zone Ⓐ), the controller works in a servo mode taking the system to the optimal operating point. Then (zone Ⓑ), the system undergoes a load which is approximately the same as that performed in scenario ④ but now in closed-loop, that is, with the controller taking action to reject the external disturbance. It is worth referring that the time scale of the graphs of Sect. 4.2 is more spread than the one used in the graphs of the open-loop study (Sect. 4.1).

All the results exhibited in this section were obtained with a predictive horizon $p = 20$, a control horizon $m = 10$, and $\Delta t = 5\,\text{min}$.

Figure 4 portrays the controller actions and their repercussions on the controlled variables.

Zone Ⓐ reveals the performance of the controller in a scenario of servo control of variables y'_{E} and $x_{\text{lt,M}}$. In order to perform this test, the values of the variables presented in Table 4 were changed at $t = 50.1\,\text{h}$ to: $y_{\text{sp},1} = 96.6\%(m/m)$, $y_{\text{sp},2} = 20.0\%(n/n)$, $u_{\text{ref},1} = 326\,\text{kg h}^{-1}$, and $w_3 = 0.01\,\text{h}^2\,\text{kg}^{-2}$.

Taking into consideration the oil flow rate that one wants to process, the controller manipulates the temperature of this feed stream, T_{O}, the output temperature of the heat exchanger, T_{H}, and the flow rate of the other reactant (methanol) that also feeds the reactor, F_{M}. Pursuing the setpoint and reference values with the minimum possible cost while fulfilling the restrictions imposed (defined by the lower and upper bounds indicated in Table 4), the controller drives the system to a new steady-state. By reducing F_{M} (top graph of Fig. 4), it pushes the controlled variables down to their setpoints (middle and bottom graphs of Fig. 4). In particular, y'_{E} gets close to the limit of its acceptable range, which is a more profitable operating point. To attenuate the effect of the reduction of the methanol to oil ratio in the extent of the reaction, the controller indirectly increases the reactor temperature at the cost of a slight increase of the pre-heating of the oil that enters the reactor (T_{O} increases, as shown by the top graph of Fig. 4). During the transient period (when the controller action is reducing drastically y'_{E} and $x_{\text{lt,M}}$), the controller increases the power extracted at the heat exchanger, that is, reduces T_{H} and thus the decanter temperature. Such reduction moves the phase equilibrium between the heavy and the light phases in the decanter. After that initial sudden decrease, T_{H} is gradually increased towards a new steady-state value which is smaller than that of the initial steady-state.

Figure 5 depicts the evolution of the state variables that describe the composition profiles of the mixtures inside the reactor and the decanter. The action of the controller in zone Ⓐ indirectly decreases the fraction of methanol in the reacting mixture and increase the molar fractions of E and G. Although several phenomena are involved in the evolution of these variables (namely the reaction extent), the considerably smaller amount of M present in the reacting mixture, makes the amounts of E and G more representative and therefore with higher molar fractions. Such effect is not perceptible for TG, DG, and MG since their amounts are insignificant when compared with other components. As revealed by the middle and the bottom graphs of Fig. 5, the action of

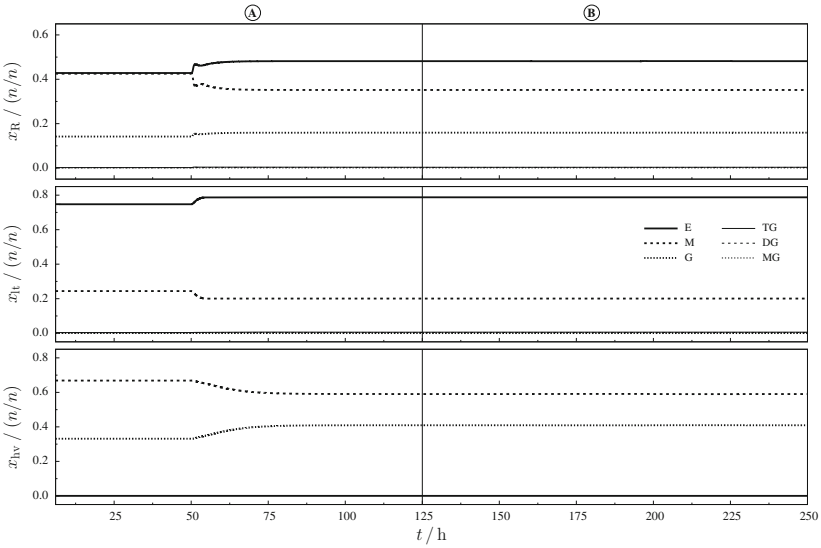


Fig. 5 Closed-loop response of the reactor-decanter system (state variables)

the controller also interferes with the composition of the streams leaving the decanter, making the upper stream richer in E, the bottom stream richer in glycerol, and both the upper and the down stream poorer in M, relatively to the original steady-state.

In zone Ⓑ, the performance of the controller is accessed by analysing its ability to handle the external disturbance $F_O = 3000 + 200\,\mathcal{H}(t - 160.1) - 400\,\mathcal{H}(t - 195.8) + 200\,\mathcal{H}(t - 225.1)$ that affects the system. The effect of such load in the system in the absence of the controller was studied in scenario ④ of Sect. 4.1. Comparing the evolution of the system in the appropriate sections of Figs. 4 and 2, it is clear that the controller effectively keeps the controlled variables in their desirable values, neutralizing the effect of the disturbance in F_O. The deviation from the steady-state that the system otherwise would suffer is completely avoided by the controller. According to the top graph of Fig. 4, such result is achieved by manipulating slightly F_M and T_H.

In what concerns the state variables that define the composition of the mixtures in the reactor and in the decanter, it is possible to see in Fig. 5 that under the corrective action of the controller they practically do not suffer any change.

5 Conclusions

The main contributions of this work are the development and a use case demonstration of a mechanistic dynamic model of a continuous biodiesel production line. In order to contemplate the main phenomena occurring in each of the sections of the

process as well as to take into account the existing interactions, the model includes a reactor, a heat exchanger, a decanter, a washer, and a dryer units. The reactor model describes the temperature and the composition of the main species involved in the transesterification reaction. The computational burden of the decanter model was contained with the use of a neural network that approximated the iterative calculations of the liquid-liquid equilibria. Finally, the washer and the dryer were modeled considering an ideal operation.

One of the presented use cases is the open-loop simulation of the production line that allows to predict the process dynamics and the final product quality. Such simulation may be of interest for what-if analysis and for off-line training of the process operation teams. This use case is followed by a closed-loop study in which a nonlinear model predictive controller carries out multivariable control and optimization of the production process taking into account the operational and quality constraints. This control approach may render tangible economic results via the minimization of the energy and the reactant specific consumption.

The model developed herein takes into account the main reactions that take place in the system. Its predictive capability may be further improved with the incorporation of the side reactions, such as the saponification due to water and high level of free fatty acids in the oil, that reduce the yield of the process in the reaction unit and hinder the phase separation in the decanter.

References

1. Benavides, P.T., Diwekar, U.: Optimal control of biodiesel production in a batch reactor: part I: deterministic control. Fuel **94**, 211–217 (2012). https://doi.org/10.1016/j.fuel.2011.08.035
2. Bock, H.G., Plitt, K.J.: A multiple shooting algorithm for direct solution of optimal control. In: Proceedings of the 9th IFAC World Congress, pp. 242–247. Pergamon Press, Budapest (1984)
3. Brásio, A.S.R., Romanenko, A., Leal, J., Santos, L.O., Fernandes, N.C.P.: Nonlinear model predictive control of biodiesel production via transesterification of used vegetable oils. J. Process Control **23**(10), 1471–1479 (2013). https://doi.org/10.1016/j.jprocont.2013.09.023. ISSN: 0959–1524
4. Brásio, A.S.R., Romanenko, A., Fernandes, N.C.P.: Development of a numerically efficient biodiesel decanter simulator. In: Oliveira, J.F., Almeida, J.P., Pinto, A.A. (eds.) CIM Series in Mathematical Sciences. Springer-Verlag, Berlin (2014)
5. Brásio, A.S.R., Romanenko, A., Santos, L.O., Fernandes, N.C.P.: First principle modeling and predictive control of a continuous biodiesel plant. J. Process Control **47**, 11–21 (2016). https://doi.org/10.1016/j.jprocont.2016.09.003. ISSN: 0959–1524
6. Charles, C., Gerasimchuk, I., Bridle, R., Moerenhout, T., Asmelash, E., Laan, T.: Biofuels — At what cost? A review of costs and benefits of EU biofuel policies. Technical report, International Institute for Sustainable Development (IISD), April 2013. https://www.iisd.org/gsi/sites/default/files/biofuels_subsidies_eu_review.pdf
7. EN 14214. European Standard EN 14214.: Automotive fuels - Fatty acid methyl esters (FAME) for diesel engine - Requirements and test methods. CEN - European Committee for Standardization, Brussels, Belgium (2008)
8. Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S.: SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. ACM T. Math. Software **31**(3), 363–396 (2005)

9. Issariyakul, T., Dalai, A.K.: Comparative kinetics of transesterification for biodiesel production from palm oil and mustard oil. Can. J. Chem. Eng. **90**(2), 342–350 (2012). https://doi.org/10.1002/cjce.20679
10. Julier, S.J., Uhlmann, J.K.: Unscented filtering and nonlinear estimation. Proc. IEEE **92**(3), 401–422 (2004). https://doi.org/10.1109/JPROC.2003.823141. ISSN: 0018–9219
11. Mjalli, F.S., Hussain, M.A.: Approximate predictive versus self-tuning adaptive control strategies of biodiesel reactors. Ind. Eng. Chem. Res. **48**(24), 11034–11047 (2009). https://doi.org/10.1021/ie900930k
12. Moser, B.R.: Biodiesel production, properties, and feedstocks. In Vitro Cell. Dev. Biol. Plant **45**(3), 229–266 (2009). Springer-Verlag. https://doi.org/10.1007/s11627-009-9204-z. ISSN: 1054–5476
13. Noureddini, H., Zhu, D.: Kinetics of transesterification of soybean oil. J. Am. Oil Chem. Soc. **74**(11), 1457–1463 (1997). https://doi.org/10.1007/s11746-997-0254-2
14. Santos, L.O., Oliveira, N.M.C., Biegler, L.T.: Reliable and Efficient Optimization Strategies for Nonlinear Model Predictive Control. In: Rawlings, J.B. (ed.) Proceedings of the DYCORD, pp. 33–38. Elsevier Science, Oxford (1995)
15. Shen, Y.H., Cheng, J.K., Ward, J.D., Yu, C.C.: Design and control of biodiesel production processes with phase split and recycle in the reactor system. J. Taiwan Inst. Chem. Eng. **42**(5), 741–750 (2011). https://doi.org/10.1016/j.jtice.2011.01.010. ISSN: 1876-1070
16. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**(1), 25–57 (2005). https://doi.org/10.1007/s10107-004-0559-y. ISSN: 1436–4646

Prior Information in Bayesian Linear Multivariate Regression

J. Casaca

Abstract The paper introduces the Bayesian approach to multivariate regression analysis, from a subjective point of view. A review of non-informative and informative priors adequate to practical situations is carried out. The marginal posteriors of the regression coefficients and the variance factors corresponding to the Laplace, Jeffreys and conjugate priors, as well as the respective modes, are presented. Of note is the fact that Laplace and Jeffreys priors, as it would be expected of non-informative priors, yield maximum posterior estimates of the regression coefficients identical to the maximum likelihood estimate.

Keywords Elicitation · Hyper-parameter · Likelihood · Posterior PDF · Prior PDF

1 Introduction

1.1 Invariance and Propriety

According to Christian Robert [13]: "..., the most critical and most criticized point of Bayesian analysis deals with the choice of the prior distribution, since, once this prior distribution is known, inference can be led in an almost mechanical way ...".

A first general requirement for a prior distribution of the parameters is that it must be invariant under transformation of variables. The most important transformations correspond to changes in the units of the variables, which belong to the general group of similarity transformations. However, sometimes other transformations of the variables are necessary, such as in the case of beta regression, gamma regression, Poisson regression etc. The necessity of objective (non-informative) invariant prior distributions has led to the concept of reference prior distribution [4, 5], which, under favourable mathematical conditions, reduces to the Jeffreys prior distribution [13].

J. Casaca (✉)
Laboratorio Nacional de Engenharia Civil, Av. do Brasil 101,
1700-066 Lisbon, Portugal
e-mail: jmmcasaca@gmail.com

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_7

A second general requirement, that the prior distribution of the parameters is represented by a proper prior probability density function (PDF) is not really critical. With the exception of the conjugate distribution families, which have proper PDF, most of the theoretical prior distributions (Laplace, Jeffreys, maximum entropy, reference etc.) are represented by improper prior PDF, that we will call prior pseudo-PDF. The really critical requirement is that the integral of the product of the likelihood by the prior pseudo-PDF is proper and, therefore, the joint posterior PDF of the parameters is proper [3].

1.2 Objectivists Versus Subjectivists

In Bayesian analysis, there are two major attitudes, affiliated in different schools of probability, towards the nature of the prior information and its representation by a prior distribution [1, 12]: (i) the objectivists consider that any prior information other than the restrictions on the mathematical models imposed by the Laws of Nature is subjective and therefore not scientific [9]; (ii) the subjectivists mostly concerned with decision in practical situations, defend the validity of any prior information on the distribution of the parameters in possession of the decision-maker [11, 14].

For an objectivist, the prior distribution of the parameters must be objective, i.e., it must respect the constraints imposed by the Laws of Nature and provide minimum information about the parameters [2, 9]. This requirement is satisfied by the maximum entropy prior distributions [9]. More recently, an effort is being made to construct reference priors that are invariant under groups of transformations and objective in the sense of being non-informative [4].

For the subjectivists, the prior distribution should convey information on the parameters and, preferably, should belong to the conjugate family of the sample distribution. The so called prior conjugate is a prior distribution such that the posterior distribution belongs to the same (conjugate) family of the prior [3]. Many sampling distributions have conjugate families (the normal-inverted gamma family is conjugate of the normal sample, the gamma family is conjugate of the Poisson sample, the beta family is conjugate of the binomial sample etc.). In this case, an additional problem, to be solved by the decision-maker, is the elicitation of the hyperparameters of the conjugate prior PDF, which may prove to be decisive in the further inferences.

1.3 The Case of Regression Analysis

In regression analysis, as a rule, the decision-maker has some general ideas on the basic functions he wants to use to model the relation between covariates and responses. The two most common states of prior knowledge regarding the regression parameters are: (i) total lack of information on the distribution of the

regression parameters; (ii) a certain amount of prior experience on modelling the relation between the same type of covariates and responses, such that the elicitation of a prior distribution to the regression parameters is possible.

To face these two distinct situations, the decision-maker needs a non-informative prior distribution and an informative prior distribution. When the regression problem lies within the frame of the Gauss–Markov model: (i) the Laplace prior distribution, since is invariant to similarity transformations and may be regarded as a mean to "normalize" the likelihood, i.e., to transform the likelihood into a posterior PDF of the parameters, may be considered as the best option; (ii) to convey prior information on the distribution of the regression parameters, the best option is the conjugate normal-inverted gamma proper prior distribution.

1.4 The Scope of the Paper

The paper is directed to the common users of multivariate linear regression analysis who are interested in upgrading their knowledge of Statistical Inference with Bayesian concepts. The author is sympathetic to the subjective point of view and, therefore, recommends the use of informative conjugate prior distributions, whenever data from prior experiences is available to elicitate the hyper-parameters of the prior PDF.

The paper introduces briefly the Bayesian approach to multivariate linear regression analysis and then presents a synthesis of the Bayesian regression parameter estimation formulae under the scenarios defined by: (i) the Laplace and Jeffreys prior pseudo-PDF; (ii) the conjugate normal-inverted gamma prior PDF.

2 Multivariate Regression Analysis

2.1 Covariates, Responses and Basic Functions

The main objective of multivariate regression analysis is to identify an analytic expression adequate to model the relation (regression) between a set of independent variables $(x_1, \ldots, x_r)$, the covariates, the regressor variables, the input variables etc., and a dependent variable (y), the response variable, the predicted variable, the output variable etc. The unknown relation $y = \phi(x_1, \ldots, x_r)$ is approximated by a linear relation [7]:

$$y = \phi(x_1, \ldots, x_r) \approx \sum_{i=1}^{n} \beta_i \varphi_i(x_1, \ldots, x_r) + \varepsilon \tag{1}$$

where: (i) the β_i are the (n) unknown regression coefficients (including an intercept); (ii) the φ_i are (n) basic functions of the covariates such as polynomial terms,

trigonometric functions, logarithms etc.; (iii) ε is a noise component which expresses simultaneously the approximation error of the model (epistemic error) and the observation error of the response (y). The covariates are supposed to be measured without significant error.

2.2 The Gauss–Markov Model

If m responses $y^\top = (y_1, \ldots, y_m)$ are observed, the m linear relations (1) may be expressed in matrix form:

$$y = B\beta + \varepsilon \tag{2}$$

where: (i) $y^\top = (y_1, \ldots, y_m)$ is the vector of the responses; (ii) $B(m, n)$ is the matrix of the regression basic functions also called the design matrix [7]; (iii) $\beta(n, 1)$ is the vector of the regression coefficients; (iv) $\varepsilon(m, 1)$ is the noise vector which results from the response's observation errors and the model's lack of adequacy (epistemic error).

If the noise vector has a Gaussian distribution with null mean vector ($E(\varepsilon) = 0$) and variance matrix of the form $V(\varepsilon) = \sigma I$, where σ is an unknown positive variance factor, and I is the identity matrix of order m, the model is said to be a homoscedastic Gauss–Markov model. In a Gauss–Markov model, the PDF of the responses vector (y) is:

$$f(y|B, \beta, \sigma) = (2\pi\sigma)^{-m/2} \exp\left(\frac{(y - B\beta)^\top (y - B\beta)}{2\sigma}\right) \tag{3}$$

Since the variance, and not the standard deviation, is the variable of interest, for a question of commodity in differentiation and integration, the symbol σ is used to represent the variance instead of the standard deviation as usual.

2.3 The Likelihood

The likelihood function of the responses vector (y) is the function with the same analytical expression of the PDF (3) but where the parameters (β, σ) become variables and the responses vector (y) becomes the parameter:

$$L(\beta, \sigma|B, y) = (2\pi\sigma)^{-m/2} \exp\left(\frac{(y - B\beta)^\top (y - B\beta)}{2\sigma}\right) \tag{4}$$

The maximum likelihood estimator of the unknown regression coefficients vector (β) is:

$$\beta_{ML} = (B^\top B)^{-1} B^\top y \tag{5}$$

The maximum likelihood estimator (σ_{ML}) and the unbiased estimator (s) of the unknown variance factor (σ) are:

$$\text{(i)}\,\sigma_{ML} = \frac{1}{m} v^{\top} v; \quad \text{(ii)}\, s = \frac{1}{m-n} v^{\top} v \tag{6}$$

where:

$$v = y - B\beta_{ML} \tag{7}$$

is the vector of the residuals from β_{ML}. The maximum likelihood estimador (σ_{ML}) is a biased estimator that underestimates the variance factor (σ). The estimator with Bessel's correction (s) is an unbiased estimator of the variance factor (σ).

2.4 Fisher's Information

The Fisher's information is a measure of the information provided by the responses vector (y) to the estimation of the regression parameters (β, σ). The Fisher's information (FI) is given by the determinant of the symmetric of the mathematical expectation of the Hessian matrix of the log-likelihood (natural logarithm of the likelihood):

$$FI(\beta, \sigma) = \frac{m}{2\sigma^{n+2}} \det(B^{\top} B) \tag{8}$$

The Fisher's information increases with the number of responses and the determinant of the precision matrix ($\sigma^{-n} B^{\top} B$) of the maximum likelihood estimator of the regression coefficients (5).

3 Bayesian Estimation of the Regression Parameters

3.1 The Formula of Bayes-Laplace

In Bayesian analysis, the unknown regression parameters (β, σ) that regulate the PDF of the observed responses (y) are regarded as the outcomes of two random variables with a joint PDF $h(\beta, \sigma)$, which is called the joint prior PDF of the parameters. A primary objective of Bayesian analysis is the estimation of the regression parameters, taking into account both the likelihood $L(\beta, \sigma | B, y)$ and the joint prior PDF of the regression parameters $h(\beta, \sigma)$.

The formula of Bayes-Laplace relates the joint posterior PDF of the parameters $p(\beta, \sigma | B, y)$ to the likelihood $L(\beta, \sigma | B, y)$ and the joint prior PDF $h(\beta, \sigma)$:

$$p(\beta, \sigma | B, y) = \frac{L(\beta, \sigma | B, y) h(\beta, \sigma)}{\int_0^{+\infty} \int_{\Re^n} L(\beta, \sigma | B, y) h(\beta, \sigma)\, d\beta d\sigma} \tag{9}$$

where the integral:

$$p(y|B) = \int_0^{+\infty} \int_{\Re^n} L(\beta, \sigma | B, y) h(\beta, \sigma)\, d\beta d\sigma \tag{10}$$

does not depend of the regression parameters (β, σ) and is constant for every given vector of responses (y). The integral (10) is called the prior predictive PDF of the responses, given the model defined by the design matrix B. The prior predictive PDF $p(y|B)$ plays a central role in the Bayesian methodology used to compare different models (Bayes factor).

3.2 The Mode of the Joint and the Marginal Posteriors

The frequentist school recommends the estimation of the regression parameters (β, σ) with the maxima $(\beta_{ML}, \sigma_{ML})$ of the likelihood (4). In a similar way, the orthodox Bayesian approach recommends the estimation of the regression parameters with the maximum joint posterior PDF estimates $(\beta_{MP}, \sigma_{MP})$.

In practice the regression parameters (β, σ) are better estimated with the marginal posterior PDF of the regression parameters. The marginal posterior PDF of the regression coefficients vector (β) is obtained by marginalization of the variance factor (σ) in the joint posterior PDF (9):

$$p(\beta | B, y) = \int_0^{+\infty} L(\beta, \sigma | B, y) h(\beta, \sigma)\, d\sigma \tag{11}$$

The marginal posterior PDF of the variance factor (σ) is obtained by marginalization of the regression coefficients vector (β) in the joint posterior PDF (9):

$$p(\sigma | B, y) = \int_{\Re^n} L(\beta, \sigma | B, y) h(\beta, \sigma)\, d\beta \tag{12}$$

The modes (maxima) of the marginal posterior PDF (11) and (12) are orthodox Bayesian estimates of the regression coefficients vector (β) and the variance factor (σ), respectively.

3.3 Prior Information

Three prior distributions were referred as the object of this paper: the Laplace, the Jeffreys and the conjugate prior distributions. The Laplace and the conjugate options are not favoured by the objectivists. The Laplace prior distribution is criticized by many objectivists mainly because there are many situations where it performs poorly (beta-regression etc.). The conjugate prior distributions are criticized by the objectivists because they are informative and, therefore, subjective. However, the fact that the prior PDF and the posterior PDF of the parameters belong to the same family of distributions brings a great formal consistency to the analysis.

In regression analysis, within the frame of the Gauss–Markov model, the Laplace distribution has a good performance: (i) is invariant with regard to similarity transformations; (ii) originates proper joint and marginal posteriors of the regression parameters; (iii) is most adequate to model selection with the Bayes factor. With regard to theoretical consistency, the Laplace prior distribution originates a joint posterior that: (i) is an improper permissible prior [4] since it is the limit of a succession of proper posteriors generated by proper uniform priors; (ii) the Laplace posterior may be regarded as the normalized likelihood, i. e., the transformation of the likelihood into a PDF (the Laplace estimates of the regression coefficients are identical to the maximum likelihood estimates).

The Jeffreys prior distribution generates joint and marginal posteriors that estimate the regression coefficients as the maximum likelihood methods. However, the Jeffreys estimates of the variance factor are biased. Although the Jeffreys prior PDF is more adequate to situations where the responses suffer non-linear transformations (beta regression etc.), there is not a decisive reason to use the Jeffreys prior in common multivariate linear regression analysis.

4 Laplace Joint and Marginal Posterior PDF

The Laplace joint prior distribution of the regression parameters (β, σ), is defined by the constant pseudo-PDF:

$$\begin{cases} h_L(\beta, \sigma) = \kappa & (\beta \in \Re^n \wedge \sigma > 0) \\ h_L(\beta, \sigma) = 0 & (\beta \in \Re^n \wedge \sigma \leq 0) \end{cases} \tag{13}$$

which is the paradigm for an improper prior PDF, because the integral:

$$\int_0^{+\infty} \int_{\Re^n} \kappa \, d\beta d\sigma \tag{14}$$

is improper. However, in Bayesian multivariate regression analysis, within the Gauss–Markov model frame, the Laplace pseudo-prior PDF may be used as a regular PDF, because the joint and the marginal PDF of the regression parameters are proper [3].

Taking $\kappa = 1$, the Laplace prior predictive PDF is given by:

$$p_L(y|B) = \frac{\Gamma(a)\sqrt{\det(B^\top B)}}{b^a\sqrt{(2\pi)^{m-n}}} \tag{15}$$

where $\Gamma(\cdot)$ stands for the gamma function [6], a is the shape parameter (18.iii) and b is the scale parameter (18.iv) of the joint posterior PDF (16).

The joint posterior $p_L(\beta, \sigma|B, y)$ of the regression parameters is a proper normal-inverted gamma PDF [8]:

$$p_L(\beta, \sigma|B, y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{Q}{2\sigma}\right)\frac{b^a}{\Gamma(a)\sigma^{a+1}}\exp\left(-\frac{v^\top v}{2\sigma}\right) \tag{16}$$

where Q is the quadratic form:

$$Q = (\beta - \beta_{ML})^\top B^\top B(\beta - \beta_{ML}) \tag{17}$$

The four parameters of the normal-inverted gamma joint posterior (16) are the mean vector μ_L , the variance matrix V_L , the shape parameter a, and the scale parameter b:

$$\text{(i)}\,\mu_L = \beta_{ML};\quad \text{(ii)}\,V_L = \sigma(B^\top B)^{-1};\quad \text{(iii)}\,a = \frac{m-n-2}{2};\quad \text{(iv)}\,b = \frac{v^\top v}{2} \tag{18}$$

The marginal posterior PDF of the regression coefficients $p_L(\beta|B, y)$ is a proper multi-t PDF with v_L degrees of freedom, mean vector μ_L (equal to the mode) and variance matrix Σ_L [13]:

$$\text{(i)}\,v_L = m-n-2;\quad \text{(ii)}\,\mu_L = \beta_{ML};\quad \text{(iii)}\,\Sigma_L = \frac{v^\top v}{m-n-4}(B^\top B)^{-1} \tag{19}$$

The Laplace maximum marginal posterior solution for the regression coefficients is equal to the maximum likelihood solution. The variance matrix of the regression coefficients is independent of the unknown variance factor (σ).

The Laplace marginal posterior PDF of the variance factor $p_L(\sigma|B, y)$ is a proper inverted gamma PDF with mode (ω_L) different from the mean value (σ_L):

$$\text{(i)}\,\omega_L = \frac{v^\top v}{m-n};\quad \text{(ii)}\,\sigma_L = \frac{v^\top v}{m-n-4}; \tag{20}$$

The maximum of the Laplace marginal posterior of the variance factor is attained at the mode ω_L which is equal to the corrected unbiased estimator of the variance

factor (s). The mean value (σ_L) of the Laplace marginal posterior of the variance factor is identical to the variance factor of the variance matrix (19.iii) of the regression coefficients estimator β_{ML}.

5 Jeffreys Joint and Marginal Posterior PDF

The Jeffreys prior pseudo-PDF is proportional to the square root of Fishers information (8):

$$\begin{cases} h_J(\beta, \sigma) \propto \sigma^{-(n+2)/2} & (\beta \in \Re^n \wedge \sigma > 0) \\ h_J(\beta, \sigma) = 0 & (\beta \in \Re^n \wedge \sigma \leq 0) \end{cases} \tag{21}$$

In Bayesian multivariate regression analysis, within the Gauss–Markov model frame, the Jeffreys pseudo-prior PDF (21) may be used as a regular PDF, because the joint and marginal PDF of the regression parameters are proper. The Jeffreys prior pseudo-PDF $h_J(\beta, \sigma)$, introduced by Harold Jeffreys [10] in view of its invariance properties, does not provide information on the regression coefficients (β) but is informative with regard to the variance factor (σ).

The Jeffreys joint posterior $p_J(\beta, \sigma|B, y)$ of the regression parameters is a proper normal-inverted gamma PDF, witch parameters are the mean vector μ_J , the variance matrix V_J , the shape parameter a_J , and the scale parameter b_J , given by:

$$\text{(i)}\, \mu_J = \mu_L = \beta_{ML}; \quad \text{(ii)}\, V_J = V_L = \sigma (B^\top B)^{-1}; \quad \text{(iii)}\, a_J = \frac{m}{2}; \quad \text{(iv)}\, b_J = b_L = \frac{v^\top v}{2} \tag{22}$$

The Jeffreys marginal posterior PDF of the regression coefficients $p_L(\beta|B, y)$ is a proper multi-t PDF with ν_J degrees of freedom, mean vector μ_J (equal to the mode vector) and variance matrix Σ_J [13]:

$$\text{(i)}\, \nu_J = m; \quad \text{(ii)}\, \mu_J = \mu_L = \beta_{ML}; \quad \text{(iii)}\, \Sigma_J = \frac{v^\top v}{m-2} (B^\top B)^{-1} \tag{23}$$

The Jeffreys maximum marginal posterior solution for the regression coefficients is equal to the Laplace solution and to the maximum likelihood solution. The variance matrix of the regression coefficients (Σ_J) is independent of the unknown variance factor (σ).

The Jeffreys marginal posterior PDF of the variance factor $p_J(\sigma|B, y)$ is a proper inverted gamma PDF with mode ω_J and mean value σ_J:

$$\text{(i)}\, \omega_J = \frac{v^\top v}{m+2}; \quad \text{(ii)}\, \sigma_J = \frac{v^\top v}{m-2} \tag{24}$$

Both the mode (ω_J) and the mean value (σ_J) of the Jeffreys marginal posterior PDF of the variance factor are biased estimators that underestimate the variance factor (σ). The mean value (σ_J) is the variance factor of the variance matrix (23.iii) of the regression coefficients estimator β_{ML}.

6 The Conjugate Prior Distribution

In Bayesian multivariate regression analysis, within the Gauss–Markov model frame, the conjugate family of the distribution of the responses is the normal-inverted gamma family. The hyper-parameters of the conjugate joint prior PDF, which must be previously elicitated, are: (i) a mean vector μ_0; (ii) a variance matrix of the form $\sigma\,\Sigma_0$; (iii) a shape hyper-parameter a_0; (iv) a scale hyper-parameter b_0.

According to the concept of conjugate family, the conjugate joint posterior PDF of the regression parameters belongs to the normal-inverted gamma family. The conjugate marginal posterior PDF of the regression coefficients $p_C(\beta|B, y)$ is a multi-t PDF with $\nu_C = 2a_0 + m$ degrees of freedom, mean vector μ_C (equal to the mode vector) and variance matrix Σ_C [7]:

$$\text{(i)}\,\mu_C = (\Sigma_0^{-1} + B^\top B)^{-1}(\Sigma_0^{-1}\mu_0 + B^\top y); \quad \text{(ii)}\,\Sigma_C = \frac{Q_C + Q_0 + 2b_0}{2a_0 + m - 2}(\Sigma_0^{-1} + B^\top B) \tag{25}$$

where Q_C and Q_0 are the quadratic forms:

$$\text{(i)}\,Q_C = (y - B\mu_C)^\top (y - B\mu_C); \quad \text{(ii)}\,Q_0 = (\mu_C - \mu_0)^\top (\Sigma_0^{-1} + B^\top B)^{-1}(\mu_C - \mu_0) \tag{26}$$

The conjugate maximum marginal posterior solution for the regression coefficients is different from the maximum likelihood, Laplace and Jeffreys solution. The variance matrix of the regression coefficients is also independent of the unknown variance factor (σ).The conjugate marginal posterior PDF of the variance factor $p_C(\sigma|B, y)$ is an inverted gamma PDF with mode ω_C and mean value σ_C:

$$\text{(i)}\,\omega_J = \frac{Q_C + Q_0 + 2b_0}{2a_0 + m + 2}; \quad \text{(ii)}\,\sigma_J = \frac{Q_C + Q_0 + 2b_0}{2a_0 + m - 2} \tag{27}$$

Both the mode (ω_C) and the mean value (σ_C) of the conjugate marginal posterior PDF of the variance factor are biased estimators of the variance factor (σ).

7 Conclusions

A review of the joint and marginal posterior PDF of the regression coefficients and the variance factor was carried out for the Laplace, Jeffreys and conjugate prior distributions. The properties of these joint and marginal posterior PDF are very interesting

since they belong to the same families of distributions: (i) the three joint posteriors belong to the normal-inverted gamma family; (ii) the three marginal posteriors of the regression coefficients belong to the multi-t family; (iii) the three marginal posteriors of the variance factor belong to the inverted gamma family. The fact that the joint and the marginal posteriors belong to known families of distributions provides closed formulae for the parameters of the posteriors which avoids the use of the Monte Carlo simulation methods [12, 13], to obtain estimates of the regression parameters. This fact also has a positive impact on the construction of credibility regions for the parameters.

When there is no prior information on the distribution of the regression parameters, the old Laplace non-informative prior pseudo-PDF is recommended because of its simplicity and intimate relation to the likelihood. This pseudo-PDF is invariant under changes of the units of covariates and responses. James Berger, an assumed objectivist, says [3] about this prior: "Although this was routinely done by Laplace, it came under severe (though unjustified) criticism because lack of invariance under transformation". Within the restricted universe of the Gauss–Markov model, the Laplace prior distribution is the best noninformative option.

Whenever there is prior information on the distribution of the regression parameters, a conjugate normal-inverted gamma prior PDF is recommended. The correspondent posterior PDF of the regression parameters (β, σ) will belong to the same family. In this case, the elicitation of the hyper-parameters $(\mu_0, \Sigma_0, a_0, b_0)$ of the conjugate prior PDF is decisive to the construction of an adequate posterior PDF. It is of note, that when the number of responses (m) and, consequently, the Fisher's information (8) is large, the influence of the likelihood supersedes the influence of the prior on the posterior.

References

1. Barnett, V.: Comparative Statistical Inference, 3rd edn. Wiley, New York (1999)
2. Berger, J.: The case for objective bayesian analysis. Bayesian Anal. **1**(3), 385–402 (2006)
3. Berger, J.: Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer, New York (2010)
4. Berger, J., Bernardo, J., Sun, D.: The formal definition of reference priors. Ann. Stat. **37**(2), 905–938 (2009)
5. Bernardo, J.M.: Bayesian Theory. Wiley, New York (2009)
6. Casaca, J.: The Gamma, Multi-gamma, Digamma and Trigamma Functions. LNEC, ICT, INCB 17 (2012)
7. Denison, D., Holmes, C., Mallick, B., Smith, A.: Bayesian Methods for Nonlinear Classification and Regression. Wiley, New York (2002)
8. Evans, M., Hastings, N., Peacock, B.: Statistical Distributions, 3rd edn. Wiley, New York (2000)
9. Jaynes, E.: Prior Probabilities. IEEE Trans. Syst. Sci. Cybern. **sec–4**(3), 227–241 (1968)
10. Jeffreys, H.: Theory of Probability, 3rd edn. Clarendon Press, Oxford (1961)
11. Lindley, D.: Understanding Uncertainty. Wiley, New York (2006)
12. Press, S.: Subjective and Objective Bayesian Statistics, 2nd edn. Wiley, New York (2003)
13. Robert, C.: The Bayesian Choice, 2nd edn. Springer, New York (2007)
14. Savage, L.: The Foundations of Statistics, 2nd edn. Dover, New York (1972)

Perceptions of True and Fair View: Effects of Professional Status and Maturity

J. A. Gonzalo-Angulo, A. M. Garvey and L. Parte

Abstract This paper examines the effects that professional status and maturity have on the understanding and perception of the true and fair view (TFV) and its True and fair override in Spain. The effects were deduced by a survey conducted on students and auditors. The results show that, while the goal of reaching the TFV is fully integrated into the Spanish accounting system, the implications of such an objective are far from what would be expected. The evidence suggests a practical rejection of the overriding aspect associated with the TFV notion in the EU Directives which is demonstrated by a preference to follow the accounting standards in all cases rather than having to choose when not to apply them in order to achieve this objective. This aversion is logical in a country whose legislation allows little room for flexibility. Finally, the study identifies a pattern of change according to the participant's professional status and maturity. It is observed that the younger and less professional participants are more concerned with obtaining the TFV than the strict coherence with the accounting standards. However, as the participants evolve according to age and professional status they prefer the TFV to be obtained by a rigorous following of the standards without having to override them, they are also more demanding for a detailed definition of TFV and are much less favourable to the imposition of fines where the TFV is not achieved by following the accounting standards.

Keywords True and fair view · Maturity · Professional status · Survey
True and fair override

J. A. Gonzalo-Angulo · A. M. Garvey
Faculty of Economics, Business Administration and Tourism,
Universidad de Alcalá, Madrid, Spain
e-mail: josea.gonzalo@uah.es

A. M. Garvey
e-mail: anne.garvey@uah.es

L. Parte (✉)
Faculty of Economics and Business Administration,
Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
e-mail: lparte@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_8

1 Introduction

This paper addresses some issues relating to the objective of True and Fair View (TFV) for financial reporting, and the overriding principle on accounting standards implied in reaching that objective. This is not a new theme but one which continues to be topical and is a subject of continual debate in accounting research (see e.g. Alexander and Eberhartinger [2]). However, here the issue is addressed to show the effect that professional status and maturity have on the understanding and perception of the participants demonstrated with an empirical study. The reason for the study is twofold, firstly, to discover how TFV is understood by professionals and non-professional users and secondly to identify if a learning curve exists in the education process. We understand that TFV is not an easy concept to assimilate and that its understanding should change through academic formation and maturity. By discovering more about the perception of TFV we can rectify the limitations in its application and improve (principally through education material) the way that the concept is transmitted to students and applied by auditors.

The TFV is an important issue which has not always received unanimous agreement among accounting standard setters, auditors, professionals and academics over the years. It has been of particular relevance over the last years with the convergence project between International Financial Reporting Standards (IFRS) and US GAAP whereby the US were extremely reluctant to accept the override principle due to the fact that US GAAP became over time a more rules based philosophy and the idea of the flexibility introduced by the TFV override caused uncertainty in a country where detailed rules were now the norm. However, Van Hulle [20] reminds us that accounting regulators are not perfect and even when rules and standards cater for the majority of circumstances they can never cover all situations in practice. Moreover, Arden [3] mentions from a court justice point of view that whenever there are rules there will always be problems that the rules cannot solve making the TFV requirement necessary.

However, the European Union (EU) accepted the override principle in the Directives and therefore it is a legal obligation for all member states. The inclusion of the override was initially promoted by the UK on its incorporation into the EU in 1971. As mentioned by Alexander and Eberhartinger ([2], p. 571), the UK considered it important to issue a new 'opinion of Counsel' (FRC, 2008) which clearly shows its support and considers the continued importance of the TFV override. Interestingly however, Livne and McNichols [12] find that UK firms invoking more costly overrides report weaker performance and have less informative financial statements and lower earnings quality. In Spain, the override principle has been invoked in only a few cases during the period of more than two decades of operation (see Cea and Vidal [4]).

Despite the initial arguments against the inclusion of the TFV override by the US, the IASB continues to include the override position in IAS 1. The term used in IAS 1 is now Fair Presentation rather than TFV but we understand that the override clause works in a similar way. It is clear however on reading IAS 1 that the override

stipulation is not the favoured practice and should only be used in exceptional circumstances. This condition of exceptionality could be considered unfortunate due to the fact that in practice it acts as a disincentive to surpass accounting standards and rules which do not clearly offer a TFV of the financial position of the entity.

It is important to mention that the study of TFV has occupied a relevant position in the accounting literature and has faced a number of criticisms due to the requirement of giving a TFV in the financial statements (see e.g. Nobes and Parker [15], Alexander and Jermakowick [1], Alexander and Eberhartinger [2]). Prior studies have explored the different attempts to give TFV a definition (Low and Koh [13], Hamilton and Ó'Hógartaigh [7], Garvey [5]), the application of TFV in practice and the understanding of the TFV by auditors (Nobes and Parker [15], Kosmala [10]), and the perceptions of TFV by different groups (Houghton [8], Low and Koh [13], Kosmala [10], Kirk [9], Garvey et al. [6]). This study updates the literature by re-examining the perception of TFV using a sample of professional and non-professional users and in a codified legal system.

The importance of the TFV concept in the codified legal systems used in Continental Europe and the differences with the common law systems used in countries like the UK or US allow us to understand the results of the study. The accounting system being examined follows the mentioned codified legal system and is unfamiliar with a flexible thought like TFV. This TFV did not come with a definition explaining how it should be achieved which makes its application more difficult in a country where professionals are accustomed to applying the law by its wording and where there is no room for interpretation. Clearly a very different form of interpreting the law to the common law system and hence the difficulty of applying a very open and flexible thought like TFV. The new procedure required a change in mentality by the accounting practice but also by educators who were instructing those future professionals.

Then, we focus on how the TFV was explained during Spanish accounting and auditing courses and on the understanding that was obtained and transferred to those in question. In reality we are observing what kind of TFV is inculcated to Spanish students and what is finally understood by those in practice. Our results and observations come from a survey distributed to undergraduate and postgraduate accounting students. In order to identify the application of TFV in practice, this survey was also addressed to auditors in public practice. By comparing the results obtained by students and auditors we appraise the evolution of the concept due to professional status and maturity from the education period to the professional period. The results clearly note differences in the understanding and perception of the principle of TFV depending on the academic and professional advancement of the participant.

In order to remind us how the TFV stands at present, we include Fig. 1 which demonstrates how financial statement preparers should act in order to obtain the TFV. The override area is the most problematic and is the part that we understand is not always understood correctly. By proceeding with the guidelines in graph 1, a TFV would be obtained as presently required. The professional judgement used in order to obtain this result however could vary from one professional to another.

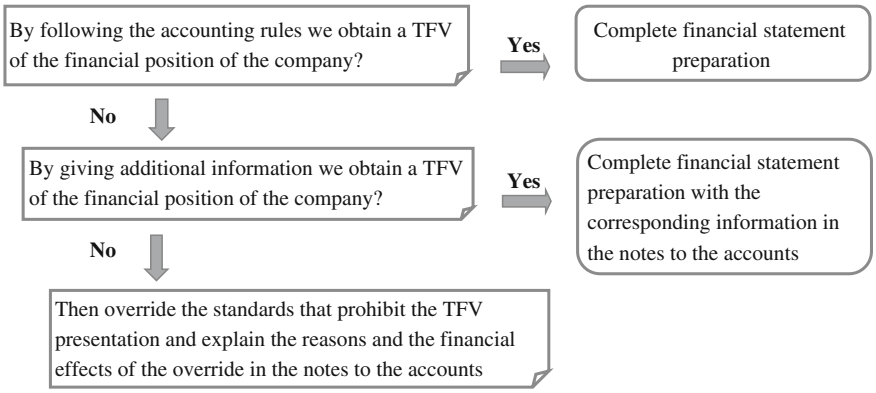


Fig. 1 Use of professional judgement required at each stage Source: Own

In this study we also focus on the importance of expertise, maturity and accounting education in relation to this complex concept. Previous research shows that professional experience has an influence on auditors' judgement (Quick and Sánchez [16]). The expert novice theory also reinforces this. The areas of age and accounting education or training are also examined and highlight differences between these groups (Montoya et al. [14]). In any field there will be a normal evolution in the learning process so perhaps that there are differences here would be naturally expected. However it is by observing the existence and location of differences that steps may be taken to eliminate them when they are not appropriate and steps to assist the learning process by enabling faster processing methods of understanding the concept correctly and thus enabling its fitting application in practice.

Our results show the complete integration of the TFV in the Spanish accounting system through the acceptance of this objective after more than twenty years of operation. Nevertheless, the version of TFV accepted in Spain could have differences with that currently accepted in the common law accounting systems of other countries. For instance, the participants show their preference for a written definition of TFV, i.e. an appeal for guidelines to help them in reaching the objective implied in TFV. On the other hand, the more mature participants bid for relief in sanctions and penalties in the case of non-compliance. The introduction of fair value (FV) is seen as positive and favourable in achieving the TFV of the company's situation and results. We detect that the effect of individual maturity, experience and accountability due to professional practice is very important in the process of opinion development of TFV.

The results obtained were not entirely surprising due to the fact that Spain forms part of the group of countries which has a codified legal base, a system whereby the law is written in a lot of detail leaving little or no room for professional judgement. For this reason we felt a necessity to question the surveyed groups about their understanding of the concept and whether they perceive that the TFV objective is achieved by simply following the accounting standards in vigour at the time of application or

whether the TFV objective requires something more. It is necessary to identify these perceptions in order to make changes for its correct application according to the law.

The remainder of the study proceeds as follows. The next section presents the literature review and hypotheses, the third section shows the research method (the survey design) and the summary statistics, the fourth section discusses the results and the final section provides our concluding remarks.

2 Literature Review and Hypotheses

This study examines the effects of professional status and maturity on the perception of TFV by observing the perception and understanding of different groups. The literature has paid special attention to the auditor's meaning and perception of TFV (Nobes and Parker [15], Kosmala [10], Kirk [9], Garvey [5], Garvey et al. [6], among others) and how the TFV is applied in practice (Nobes and Parker [15]). More interestingly, previous studies that have analysed the different perceptions of TFV between groups find mixed results. For example, Houghton [8] finds that accountants and shareholders do not share the same meaning of TFV nor do they share similar cognitive structures. Kirk [9] finds that the three group's surveyed (auditors, financial directors and shareholders of listed companies in New Zealand) share similar perceptions of the TFV; but perceive TFV to be quite different from 'fairly presents' and 'fair presentation'. In contrast, Low and Koh [13], Laswad [11] and Kosmala [10] generally do not find differences across groups.

The influence of professional experience on auditors' judgment is a widely analysed question in the audit field. Quick and Sánchez [16] examine the effect of management explanation on the auditor decision process in analytical procedures. They include the variable auditor's experience in the study because prior evidence shows that the results may be conditioned due to the level of auditor's expertise. That is, the problem solving methods used by expert auditors differs from that of more novice auditors. However, the results do not confirm previous evidence in the field. Montoya et al. [14] examine the application of materiality (permissive, moderate and strict) on the financial information using a sample of 338 Spanish auditors. The results show that in practice auditors apply different levels of materiality. Furthermore, they detect that variables such as age, academic training, firm turnover, and number and kind of companies audited, influence the effective use of the qualitative side of materiality.

Based on this literature, this study investigates the effects of professional status on the perception of the TFV in two groups of interest: accounting students and auditors. We believe this study is important because it gives an insight into the perception of this complex accounting concept at different levels of professional status and maturity which will allow standard setters' and educators' to focus on the ways to ensure that the override condition is enforced and applied correctly in the future.

In order to define the hypothesis, we follow Kirk [9] who formulated the proposition in terms of a no difference perception in the concept of TFV using three groups of participants (auditors, financial directors and shareholders of listed companies)

and three areas of interest (the meaning of the concept in financial reporting, compliance with GAAP and the law, and the requirement for financial reporting). We extend this work examining the opinion of professionals and non-professional users in four areas of interest or sub hypothesis: (i) the level of integration of the TFV concept in the Spanish accounting process, (ii) the degree of distinction between the strict compliance with accounting rules and the fulfilment of TFV, (iii) the need for a written definition of TFV, and (iv) the relationship between the use of fair value measurements and the achievement of the TFV.

If there are differences between both groups, the problem solving theory and prior empirical studies in accounting and auditing (see e.g. Houghton [8], Kirk [9], Montoya et al. [14], Hamilton and Ó'Hógartaigh [7]) can help to understand the results. The problem solving theory explains that novices and experts deal with different strategies when they have to solve a problem. Specifically, novices use a means-end analysis suggesting that when novices have to solve a new problem they need to process all the information for the first time because they do not have structured schemes to apply and solve the problem. In contrast, experts have structured schemes which they have built up through past experience. When experts have to deal with a new problem, they are able to apply the schemes to solve the new problem (see e.g. Sweller [19]). More specifically, prior empirical studies in accounting and audit have found differences in the TFV meaning and interpretation when different groups are interviewed and also when each group has different characteristics in terms of experience, age, maturity, academic training, etc. (see e.g. Houghton [8], Kirk [9], Montoya et al. [14]). Although these previous studies deal with specific context and audit dilemmas to test their hypothesis, we provide an explanation for our sample represented by professional and non-professional participants as well as our four areas of interest or sub hypothesis.

Then our first hypothesis is formulated as follows:

H1: There is no difference in the perception of the meaning of the concept 'True and Fair View' due to professional status.

The survey gathered the age information and the education level frequently used in archival research to consider conditioning effects of financial reporting practices. For example, Rankin et al. ([17], p. 372) argue that 'mature students and school leavers are likely to differ in terms of preferences, attitudes towards study and self-regulation'. The previous qualifications studied were considered to possibly affect the comprehension of the TFV concept for students.

The age and maturity of the auditors could also affect their professional judgment. The acquisition of an accounting *habitus* through a process of training in the company and the inculcation of the culture of the firm could play an important role. Rich et al. ([18], p. 105) cite several studies which document that auditor judgment is shaped by specific prior experiences of the audit process, of client misstatement and persuasion and of the client or industry. Hamilton and Ó'Hógartaigh ([7], p. 916) explain that auditors are social agents in the accounting field, whose *habitus* was formed and acquired through the process of inculcation during their educational and training period'.

The difference in opinion between students and auditors is interesting because we can observe the effect of the academic training received by this latter group in order to obtain the title of auditor. As Hamilton and Ó'Hógartaigh ([7]: 916–917) outline, in order to become an auditor an individual must undergo specific training and education which is very often controlled by the accounting body to which they wish to join. Accounting regulators can therefore guide the attitudes of future auditors through the contents of the courses and the linguistics acquired by students. It takes many years for an auditor to be admitted as a partner in an audit firm, during which time they adapt to the firm's culture. In these cases to comment on TFV not only depends on technical knowledge but also on adherence to that culture which has been inculcated over the years.

Consequently, our second hypothesis is formulated as follows:

H2: There is no difference in the perception of the concept 'True and Fair View' due to the degree of maturity and accounting education of the participants.

3 Research Method

3.1 Survey Design

The survey consisted of 14 closed-form questions. The questions used a 5 point Likert attitudinal scale, ranging from 'strongly agree' to 'strongly disagree', or 'excellent' to 'poor'. The survey was pretested on a number of academics for face validity and content. To capture the perceptions of the meaning of TFV between the groups and the degree of maturity, four sections are defined. The first section deals with the level of integration of the TFV concept in the Spanish accounting process (questions 1 and 2). The second section deals with the degree of distinction between the strict compliance with accounting rules and the fulfilment of TFV (questions 3–7). The third section investigates the need for a written definition of TFV (questions 8–11). Finally, the forth section examines the relationship between the use of fair value measurements and the achievement of the TFV (questions 12–14).

To capture perceptions of TFV emphasized in the course of accounting degrees, the sample was drawn from various universities (undergraduate students) and an audit master course (postgraduate students). In particular, the Auditing School of the ICJCE (Instituto de Censores Jurados de Cuentas de España, the most important auditor body) passed the survey to the postgraduates attending the tests for the preparation of the Master in Auditing (distance learning program) at Alcalá University. This Master Degree is followed by students as the first step of the exam to attaining the qualification of auditor. The second and final step is a practical exam to be done after the end of the practical experience period of three years. All postgraduates were enrolled in the major auditing master course to access the auditor profession in Spain. The participants of the master came from different regions of Spain, thus avoiding bias due to geographical location.

In the case of the auditors we obtained help from the AECA (Asociación Española de Contabilidad y Administración de Empresas) who distributed the survey by email to its members in audit practice. The members of AECA represent the principle auditing firms in Spain, including the BIG 4 companies.

Furthermore, the profile of undergraduates is different from postgraduates because the first are merely students but in the more advanced years of their studies and the second are normally employees of auditors in public practice or audit firms. So, we have three very different groups: pure students (undergraduate); students with some experience in performing audits but without any responsibility to give an opinion (postgraduate) and auditors fully engaged in performing audits and writing audit reports.

The surveys were administered to classes between 2006 and 2008. It is noted that there was a change in the legislation in 2007 but this should not affect our participants opinions as they all began studies under the previous requirements and the new additions in this area are not fundamental but more to clarify how to achieve TFV. The students filled out a paper version of the survey that was placed on their chairs. We used this approach in an attempt to obtain a large response rate. A total of 324 usable responses were obtained.

3.2 Summary Statistics and Data Issues

Table 1 shows the sample characteristics. Of the 414 respondents, the position title of respondents included students ($n = 324$), and auditors ($n = 90$). In terms of students' level of education, 54.63% were undergraduate and 45.37% postgraduate. Table 1 also provides the age of participants. The respondents present ages between 17 and 25 ($n = 200$), between 25 and 45 ($n = 164$) and more than 45 ($n = 50$).

Table 1 Sample characteristics

Degree		Obs.
Students	Undergraduates	177
	Postgraduates	147
Auditors		90
Total		414
Age		Obs.
Between 17–25		200
Between 25–45		164
More than 45		50
Total		414

4 Results and Discussions

4.1 Univariate Analysis

4.1.1 The Integration of the True and Fair View Concept in the Spanish Accounting System

We conducted a set of questions that ask the participants if they consider the TFV concept to be an integrated concept in the Spanish accounting process after more than twenty years of mandatory application in companies accounting and auditing. In this Section, we consider two questions and we predict that the participants endorse the integration of TFV in Spanish accounting legislation:

The integration of the True and Fair View concept in the Spanish accounting system	Expected Results
Q1. The true and fair view concept is foreign to Spanish accounting	*Disagree*
Q2. The true and fair view should be abandoned in European accounting due to the fact that it is not important	*Disagree*

Table 2, panel A shows that 78.08% (80.00%) of the students (auditors) surveyed consider that the TFV is not foreign to Spanish accounting and 84.56% (92.22%) consider that the TFV concept should not be abandoned in European accounting. Both groups recognise the TFV as being part of Spanish accounting, the percentage is slightly higher in the case of professionals, the same tendency is observed in the case of whether TFV should be abandoned due to its lack of importance. In the latter case, the difference is substantially higher in the case of professionals even when both groups give a clear response to this question. We understand that these higher rates in the case of professionals is due to them working with the concept on a daily basis whereas students are applying what they have learnt in the lecture hall. It is noted that the averages do not exceed the value of 1.83 and are significantly different from 3. Table 2, panel B shows that there are no statistically significant differences depending on the maturity of participants in these questions ($p > 0.05$). The correlation of Pearson and Spearman (not reported) between both questions are statically significant ($p < 0.01$). In sum, the evidence shows that the TFV is an integrated concept in Spanish accounting legislation and there is no difference in terms of professional status, maturity and accounting education. The two hypotheses (H1 and H2) cannot be rejected in the area of interest of this Section.

As the TFV concept is incorporated into European legislation, it is obvious that it forms part of Spanish legislation from the moment of the incorporation of Spain into the European Union, and more precisely from the moment its national laws were adapted to those of the European Union. The 1990 General Accounting Plan and the reform of the accounting plan in 2007 are some of the attempts that comply with this adaptation.

Table 2 The true and fair view is an integrated concept in the Spanish accounting system

Panel A: Survey responses conditional on professional status

	Students (1)					Auditors (2)					Differ. (1)–(2) p-value	
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	t-mean	U-MannW
Q1	324	13.58	78.08	1.83***	1.18	90	16.67	80.00	1.71***	1.19	0.39	0.13
Q2	324	7.71	84.56	1.60***	0.99	90	1.11	92.22	1.40***	0.67	0.02	0.26

Panel B: Survey responses conditional on maturity and accounting education of participants

	Age (17–25)					Age (25–45)					Age (+45)					(1–2–3) Kruskal Wallis
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	p-value
Q1	200	11.00	80.00	1.76***	1.10	164	15.85	77.44	1.86***	1.23	50	22.00	76.00	1.82***	1.34	0.74
Q2	200	5.50	88.50	1.51***	0.87	164	8.54	81.71	1.67***	1.04	50	2.00	92.00	1.38***	0.70	0.21

The table reports summary statistics on the representativeness of both the professional status (Panel A) and the maturity of participants (Panel B). The panels show the percentages of agreement (4 or 5 in the Likert scale) and disagreement (1 or 2 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column of each panel shows the *p-value* calculated from the t-tests and U-Mann Whitney test (Panel A) and Kruskal–Wallis test (Panel B)

The set of answers suggests that the TFV objective is definitively adopted and forms part of the professional mentality of both, students and auditors, and that this principle of accounting has been accepted as part of the Spanish accounting system after almost twenty years of its incorporation into the Law. A possible interpretation of this is that the former codified legal based accounting system is now already transformed, thus acquiring features from common law systems converting it into a mixed system in some ways.

4.1.2 Distinction Between the Strict Compliance with Accounting Rules and the Fulfilment of True and Fair View

The next set of questions explores the distinction between the strict compliance with accounting rules and the fulfilment of TFV. Specifically, we consider five questions:

Distinction between the strict compliance with accounting rules and the fulfilment of True and Fair View	Expected results
Q3. The true and fair view is always obtained by following the accounting standards	*Disagree*
Q4. The true and fair view should sometimes include more than the compliance with the accounting standards in vigour	*Agree*
Q5. In the case where more than one true and fair view can be obtained, the one that is closer to the accounting standards is the one that should prevail	*Disagree*
Q6. In reality, the accounting standards would have to be abandoned only in exceptional cases in order to show a true and fair view	*Agree*
Q7. The non-compliance with the true and fair view in cases where the accounting standards have been strictly complied with should not be subject to a fine	*Disagree*

Question Q3 was expressly included to investigate whether there is an understanding in Spain to override one or several accounting standards in order to show a TFV, if by using them the TFV objective is not achieved. We expected that participants would be in disagreement with the question because the TFV is not always reached by strictly following the accounting standards.

Table 3, panel A, shows that 40.74% of students disagree (moderately or strongly) and 47.53% agree (moderately or strongly) and that the mean is not statistically different from 3. The evidence could be interpreted that there is not a prevalent view among students on the way the accounting standards could operate in order to draw the appropriate picture of transactions and other economic events through financial reporting. This result is achieved at the expense of students knowing that the accounting standards provide alternatives for the recognition of accounting events.

Table 3 The compliance of true and fair view

Panel A: Survey responses conditional on professional status

	Students (1)					Auditors (2)					Differ. (1)–(2) p-value	
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	t-mean	U-MannW
Q3	324	47.53	40.74	3.09	1.13	90	66.67	30.00	3.38***	1.24	0.04	0.02
Q4	324	89.82	3.71	4.20***	0.76	90	91.11	1.11	4.34***	0.67	0.11	0.13
Q5	324	63.89	15.12	3.60***	1.06	90	78.89	11.11	3.93***	1.10	0.01	0.00
Q6	324	58.34	26.24	3.37***	1.26	90	66.67	25.56	3.54***	1.33	0.25	0.13
Q7	324	24.69	53.09	2.57***	1.21	90	54.44	31.11	3.30**	1.36	0.00	0.00

Panel B: Survey responses conditional on maturity and accounting education of participants

	Age (17–25)					Age (25–45)					Age (+45)					(1–2–3) Kruskal Wallis
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	p-value
Q3	200	46.00	41.00	3.08	1.12	164	53.05	38.41	3.16	1.15	50	70.00	28.00	3.44**	1.31	0.08
Q4	200	90.50	3.50	4.22***	0.75	164	89.02	3.05	4.22***	0.74	50	92.00	8.00	4.34***	0.69	0.59
Q5	200	63.50	13.00	3.62***	1.03	164	67.68	15.85	3.63***	1.07	50	80.00	14.00	4.00***	1.21	0.01
Q6	200	60.00	24.00	3.43***	1.21	164	57.32	28.66	3.32***	1.29	50	70.00	26.00	3.62***	1.46	0.17
Q7	200	21.00	58.00	2.46***	1.16	164	35.37	43.29	2.84***	1.24	50	58.00	26.00	3.48**	1.49	0.00

The table reports summary statistics on the representativeness of both the professional status (Panel A) and the maturity of participants (Panel B). The panels show the percentages of agreement (4 or 5 in the Likert scale) and disagreement (1 or 2 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column of each panel shows the *p-value* calculated from the t-tests and U-Mann Whitney test (Panel A) and Kruskal–Wallis test (Panel B)

The results obtained from auditor's responses present a different perspective on this issue: they are more comfortable following the accounting standards to give the financial picture of companies. Table 3, panel A, shows that 30.00% of auditors disagree (moderately or strongly) and 66.67% agree (moderately or strongly). There are statistically significant differences depending on the professional status of participants ($p < 0.05$ in mean and median) and on the maturity of the participants ($p = 0.08$) in this question (Table 3, panel B). Interestingly, the gap between the agreement and disagreement for this item is higher in the more mature group, showing that in line with their maturity, their opinion on the validity of accounting standards as the way to reach TFV is stronger.

Question Q4 is related to the TFV in that it should sometimes include more than the compliance with the accounting standards in force. Our result shows that 89.82% of the students agree with the question and 91.11% of the auditors agree with the question. No differences exist relating to the group and age of the participants. Also, participants agree (moderately or strongly) with question Q5 that in the case where more than one TFV can be obtained, the one that is closer to the accounting standards is the one that should prevail (63.89% for students, and 78.89% for auditors). Interestingly, differences exist in relation to the professional level and maturity of the participant ($p < 0.05$).

The set of responses to questions Q4 and Q5 suggest that future professionals and auditors want to search for solutions to TFV problems by taking inspiration from actual accounting standards. We expect that the best way of showing the TFV is not always the one that complies more adequately with the accounting standards and it would be necessary to examine each individual case. For this reason, the desired answer would not be in agreement with this declaration but with the opposite (that each case should be examined individually, because the best way of showing the TFV is not necessarily the one that is closest to the accounting standards). The answer is not surprising however given a more rational professional behaviour, which would be to keep to the most secure, which is to follow closely the standards in force at the time of the preparation of the financial statements. This is demonstrated by the results obtained in question Q3.

Moreover, question Q6 suggests that the accounting standards would have to be abandoned only in exceptional cases in order to show a TFV reaches an agreement of 58.34% for students and 66.67% for auditors. Although agreement with the question is higher for auditors than students, the differences are not statistically significant.

The reply obtained here ties in with the legislation of the European Union and the Spanish Accounting Plan in force at the time of the survey. However, this answer differs to the one observed in question Q3, where students and auditors surveyed agree, with a high percentage that the TFV is always achieved by following the accounting standards.

Finally, question Q7 was included to observe the state of opinion on the enforcement of the TFV that in some situations implies penalties, compensations or fines to companies or auditors in the case of non-compliance with the objective of the TFV. In this case the students disagree (53.90% that with the non-compliance of the TFV in cases where the accounting standards have been strictly complied with should not

be subject to a fine). The opposite response is found in the auditors' sample (31.11% of disagreement and 54.44% of agreement). In this question we find statistical differences between participants depending on their professional status ($p < 0.00$) and also according to their maturity ($p < 0.00$).

The intended reply here would be in disagreement with the question given that if the financial information does not show a TFV it does not comply with the objective of the law even when it follows the accounting standards strictly. This means that there is an error which can provoke negative consequences for the users of the financial statements and should therefore be fined.

On analysing the case further, it could be investigated if the fact of not showing a TFV was provoked through bad faith or not. However, a company that conforms to the accounting standards but does not obtain a TFV of the financial accounts does not comply with the legal obligations corresponding to TFV. If a fine is not imposed in this case then one could ask what the purpose of the overriding clause is in the law. No company would take the risk of overriding the accounting standards to achieve the TFV knowing that there would be no legal consequences of simply complying with the standards.

The influence of maturity and accounting education are once again important, as the Kruskal–Wallis test shows. The mean and the median figures tend to increase towards the agreement with the declaration of lack of accountability in the case of non-compliance as long as the participants age increases. We interpret this as a product of the knowledge and experience in the (sometimes difficult) application of accounting standards, as well as a form of self-protection against corrective measures by the Spanish market authorities (in the case of companies) or by the Spanish government (in the case of auditors).

The two hypotheses can be rejected because there are differences in the opinion of compliance with accounting rules due to professional status, maturity and accounting education. This part of the study which observes the distinction between the strict compliance with accounting rules and the fulfilment of TFV gives some important insights into the perceptions of students and auditors in the area. Auditors reflected that a TFV is always obtained by following the accounting standards. The students answer was not so clear here. This is an unexpected result for us taking into account the override provision of TFV. In the other questions however it shows that all participants at varying levels understand the override condition but the most surprising here is that auditors reject the possibility of fines for non-compliance with TFV where the accounting standards have been strictly adhered to.

4.1.3 A Need for a Written Definition of True and Fair View

The next set of questions explores the need for a written definition of TFV. There is not a consensus on the precise definition and scope of the TFV objective, however in a country where it is usual to have written laws and standards it could be desirable to have a formal delimitation of TFV by means of a written definition. The profile of the four questions related to the definition was designed to identify the opinion on

the advantages the participants perceive in having a wording with the description of the TFV.

A need for a written definition of True and Fair View	Expected results
Q8. A detailed definition of the true and fair view would take away from the efficiency in its application.	*Agree*
Q9. The creation of a definition of the true and fair view in relation to the annual accounts is a very difficult task.	*Agree*
Q10. The true and fair view has an absolute quality that makes it unnecessary to define.	*Agree*
Q11. It is necessary to have a definition of the true and fair view.	*Disagree*

The coefficients of Pearson and Spearman (not reported) between the four questions (Q8-Q11) are statically significant ($p < 0.05$). Question Q8 explores whether a detailed definition of the TFV would take away from the efficiency in its application. The expected answer here would be that the surveyed participants were in agreement with the question given that the TFV is a concept that can vary due to socio-economic and environmental changes, and a definition could constrict it too much. However, the hypothesis was not fulfilled in this question for the student subsample. Table 4, panel A, shows that 30.86% of students agree (moderately or strongly), 37.66% disagree (moderately or strongly) and 31.48% are indifferent. We observe a high dispersion in the responses to this question. In contrast, 25.56% of the auditors agree (moderately or strongly) and 58.89% disagree (moderately or strongly). Again, we find statistical differences depending on the professional status of participants and also by the maturity of the participants.

Once again, the interpretation of the results is that knowledge and experience determine the opinion of the group of auditors in favour of a written delimitation for the TFV objective of financial reporting. The components of this group are more comfortable with a detailed definition, because this could be helpful in performing their job. Considering the responses to the following question, it is clear that most auditors realised the difficulty of obtaining that desire.

Question Q9 looks into whether the creation of a definition of the TFV in relation to the annual accounts is a very difficult task. As expected, 56.79% of students and 60.00% of auditors believe that it is difficult to create a definition of the TFV. It is noted that the percentage of agreement is not high.

Question Q10 investigates if the TFV has an absolute quality that makes it unnecessary to define. We expected that participants would agree with the question. However, just 21.92% (30.00%) of students (auditors) agree (moderately or strongly) and 50.61% (50.00%) of students (auditors) disagree (moderately or strongly). No differences are found between groups and age of participants. After examining the answer to this question it is evident that the underlying comprehension of the concept is low and again the desire by the surveyed population to have a definition of the concept that ties into the legal system in Spain.

Finally, question Q11 explores the need to have a definition of the TFV. We expected participants to disagree with the question. That is, the reasoning for arriving

Table 4 A need for a written definition of true and fair view

Panel A: Survey responses conditional on professional status

	Students (1)					Auditors (2)					Differ. (1)–(2) p-value	
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	t-mean	U-MannW
Q8	324	30.86	37.66	2.87 **	1.10	90	25.56	58.89	2.61 ***	1.16	0.05	0.02
Q9	324	56.79	21.60	3.48***	1.09	90	60.00	23.33	3.49***	1.23	0.92	0.67
Q10	324	21.92	50.61	2.64 ***	1.07	90	30.00	50.00	2.68 **	1.24	0.80	0.91
Q11	324	67.90	14.82	3.73***	1.04	90	61.11	12.22	3.59***	1.02	0.26	0.18

Panel B: Survey responses conditional on maturity and accounting education of participants

	Age (17–25)					Age (25–45)					Age (+45)					(1–2–3) Kruskal Wallis
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	p-value
Q8	200	27.50	40.00	2.80***	1.06	164	33.54	37.80	2.93***	1.12	50	26.00	66.00	2.52**	1.30	0.03
Q9	200	55.00	21.50	3.45***	1.10	164	59.76	21.34	3.51***	1.08	50	60.00	26.00	3.50**	1.36	0.66
Q10	200	22.00	47.50	2.69***	1.07	164	23.78	53.66	2.62***	1.09	50	30.00	52.00	2.62*	1.34	0.79
Q11	200	68.50	14.00	3.76***	1.01	164	64.02	15.85	3.64***	1.06	50	66.00	10.00	3.64***	1.05	0.44

The table reports summary statistics on the representativeness of both the professional status (Panel A) and the maturity of participants (Panel B). The panels show the percentages of agreement (4 or 5 in the Likert scale) and disagreement (1 or 2 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column of each panel shows the *p-value* calculated from the t-tests and U-Mann Whitney test (Panel A) and Kruskal–Wallis test (Panel B)

at this hypothesis is the following: we understand that it is not possible to obtain an adequate definition of this concept due to its open and flexible nature, and that it can vary when there are techno-economic changes. For these reasons a definition would take away from its flexibility and would make it less necessary to use professional judgement to achieve a TFV. A possibility would be to consider guidelines that would help to achieve a TFV of the financial information but a strict definition would not operate correctly in this case in our opinion.

However, the results of the survey clearly show that most respondents believe it is necessary to have a definition of the TFV (more than 60% of agreement). In contrast, only around 12% of respondents disagreed with the question. No differences are found depending on professional status or on the maturity of participants.

In summary, the groups participating in the survey show a clear preference for a written definition of TFV even knowing the difficulty of this task. The group of auditors declares that having a detailed definition would help to reach efficiency in the application of the TFV objective in the financial reporting system.

4.1.4 The Use of Fair Value to Achieve the True and Fair View

In this section, we explore the understanding of Fair value and its relationship with the TFV. After the introduction of the fair value procedures for Spanish listed companies in applying IFRS from 2005, a broad discussion on the convenience of introducing fair value in the General Accounting Plan followed during the years of the elaboration of the 2007 reform. Finally, the fair value measures were introduced into the Plan, but only for the cases where the IFRS's do not allow for another accounting or reporting option. Thus, most financial instruments, business combinations and items initial recognition are measured at fair value. The three questions related to fair value derived from the adoption of principles close to IFRS's and TFV are as follows:

The use of Fair Value to achieve the True and Fair View	Expected results
Q12. The annual accounts prepared according to fair value can show a true and fair view.	*Agree*
Q13. The financial accounts prepared according to fair value show better the true and fair view than those prepared according to historic cost and other methods.	*Disagree*
Q14. With the introduction of IFRS (IASB), the true and fair view has less importance.	*Disagree*

For question Q12, Table 5, panel A, shows that 66.98% (84.44%) of students (auditors) agree (moderately or strongly) that the annual accounts prepared according to fair value can also show a TFV. Interestingly, we find statistical differences between both groups. The agreement of auditors is higher than the agreement obtained by students.

Question Q13 asks for a comparison between historic cost procedures and those from standards based on fair value according to IFRS's. Around 50.31% of the

Table 5 The use of fair value to achieve the true and fair view

Panel A: Survey responses conditional on professional status

	Students (1)					Auditors (2)					Differ. (1)–(2) p-value	
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	t-mean	U-MannW
Q12	324	66.98	6.48	3.77***	0.88	90	84.44	12.22	3.96***	0.95	0.08	0.01
Q13	324	50.31	11.73	3.52***	0.95	90	57.78	22.22	3.47***	1.12	0.69	0.93
Q14	324	16.97	48.77	2.49***	1.10	90	11.11	70.00	2.04***	1.07	0.00	0.00

Panel B: Survey responses conditional on maturity and accounting education of participants

	Age (17–25)					Age (25–45)					Age (+45)					(1–2–3) Kruskal Wallis
	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	N	%agr	%dis	Mean	SD	p-value
Q12	200	67.00	4.00	3.82***	0.84	164	68.29	13.41	3.68***	0.97	50	94.00	4.00	4.20***	0.76	0.00
Q13	200	54.50	5.50	3.66***	0.84	164	44.51	23.17	3.28***	1.07	50	66.00	18.00	3.64***	1.14	0.00
Q14	200	12.00	52.50	2.38***	1.04	164	21.34	48.17	2.57***	1.12	50	12.00	74.00	1.90***	1.16	0.00

The table reports summary statistics on the representativeness of both the professional status (Panel A) and the maturity of participants (Panel B). The panels show the percentages of agreement (4 or 5 in the Likert scale) and disagreement (1 or 2 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column of each panel shows the *p-value* calculated from the t-tests and U-Mann Whitney test (Panel A) and Kruskal–Wallis test (Panel B)

students agree (moderately or strongly) with the fact that financial accounts prepared according to fair value show more adequately the TFV than those prepared according to historic cost and other methods. In contrast, a strong difference is found between students and auditors when they are asked if the introduction of IFRS (IASB) implies that the TFV has less importance (question Q14). Auditors reach a high level of disagreement (70.00%) compared to students (48.77%). The hypotheses H1 and H2 can be rejected in the area of interest of this Section because there are differences of opinion from the adoption of principles close to IFRS's and TFV due to professional status, maturity and accounting education.

4.2 The Multivariate Analysis

In an attempt to reinforce the previous results, the following part of Sect. 4 of the paper outlines the findings for the variables in the model using a logistic regression and a multinomial logistic regression. Our objective is to examine if the differences found between the different perceptions of TFV can be used to predict the probability of belonging to a specific group. To do that, we chose only one question from each section, the Pearson and Spearman correlation show statistical correlations between the questions from each section and thus the question chosen is intended to have most of the information content of the section. The exception is provided by Sect. 2 called *Compliance* because question 3 related to whether there is an understanding in Spain to override one or several accounting standards in order to show a TFV is not associated with question 7 related to the enforcement of the TFV that in some situations implies penalties, compensations or fines to companies or auditors in the case of non-compliance with the objective of the TFV.

The estimation of the models takes the following expression:

$$\begin{aligned} p(x) = {} & \alpha_0 + \alpha_1 TFV\text{-}integrated\,(question\,Q2) + \\ & + \alpha_2 Compliance\text{-}always(question\,Q3) + \\ & + \alpha_3 Compliance\text{-}non\text{-}fined(question\,Q7) + \\ & + \alpha_4 Definition(question\,Q8) + \\ & + \alpha_5 RelationTFVandFV(question\,Q14) + e_t \end{aligned}$$

In the logistic regression, the dependent variable is professional status. In the multinomial regression, the dependent variable is the age of participants (proxy for maturity and accounting education). The α's are the coefficients of the independent variables in the regression model. The independent variables are the same for the logistic regression and multinomial regression.

Table 6 (Panels A and B) displays the results of the logistic regression comprising the main regression coefficients, level of significance, and odds ratio for the independent variables. The odds ratio $Exp(\alpha)$ predict the change for a unit when increasing in one independent variable, holding other variables constant.

Table 6 Results of the logistic models

	Panel A: Logistic model				Panel B: Multinomial model			
	Auditors and Students		Auditors and Postgrad		Maturity and Accounting education			
χ^2	57.95		47.60		39.93			
Sig.	0.00		0.00		0.00			
R^2 Cox y Snell	0.13		0.18		0.11			
R^2 Nagelkerke	0.21		0.25		0.05			
Correctly classified	81.13		69.66		51.00			
	Exp(B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.	Exp(B)	Sig.
c	0.125	0.001	1.052	0.943		0.695		0.024
Q2: abandoned	0.826	0.254	0.700	0.046	0.960	0.726	0.843	0.432
Q3: always	1.413	0.004	1.369	0.016	1.021	0.819	1.308	0.070
Q7: fine	1.687	0.000	1.537	0.000	1.090	0.316	1.727	0.000
Q8: efficiency	0.808	0.072	0.666	0.003	0.961	0.678	0.773	0.086
Q14: important	0.616	0.000	0.615	0.001	0.953	0.627	0.606	0.004

The table presents estimation results from a logistic regression model. Panel A presents the results from the logistic regression where the dependent variable is the professional status. In the first column, the dependent variable takes the value '1' for auditors and '0' for students. In the second column, the dependent variable takes the value '1' for auditors and '0' for postgraduates. Panel B presents the results from the multinomial regression where the dependent variable is the maturity of the participants. The variable takes the value 1, 2 or 3 for the range between 17–25, between 25–45, and more than 45, respectively. As the reference group is the youngest group (age between 17–25), the first column represents the position of middle maturity respect to the youngest group and the last column represents the older group respect to the youngest group

The results of the models are acceptable looking at the chi squared significance ($p < 0.00$) and the percentage of prediction (around 70%). The first estimation of the logit model (Panel A, column 1) represents the probability of a participant being an auditor or a student. That is the dependent variable of the logistic regression model is the 'group of reference' which adopted the value '1' for auditors and '0' for students.

The results of Table 6 (Panel A, first column) show that there is no significant difference ($p > 0.05$) between auditors and students for the question related to *the integrated concept of TFV in the accounting standards (Q2)*. The evidence suggests that both share similar behavioural beliefs in terms of the TFV as an integrated concept in Spain from the EU Directives.

There is statistical significant difference ($p < 0.01$) between auditors and students for the rest of the variables included in the regression model. The results indicate that the questions relating to *whether the TFV is always obtained by following the accounting standards (Q3)* and *the non-compliance with the TFV in cases where the accounting standards have been strictly complied with should not be subject to a fine (Q7)* are more likely to influence the auditor group (odds ratio Exp(α2) 1.431, Exp(α3) 1.687, respectively, Table 6, Panel A). That is, higher levels relating to the compliance of TFV are presented for auditors compared to students. The findings

support the univariate analysis as the percentage of agreement (and mean) in these questions are higher for auditors than for students.

Also, there is statistical significant differences ($p < 0.10$) between both groups for the questions related to *Q8. A detailed definition of the TFV would take away from the efficiency in its application.* In this question it is more likely to influence students (odds ratio Exp(α4) 0.808, Table 6, Panel A). The findings support the univariate analysis as the percentage of agreement (and mean) in these questions is higher for students than for auditors. Finally, the question related to if *the TFV has less importance with the introduction of IFRS (IASB)* also presents differences between both groups ($p < 0.01$).

The second estimation of the logit model (Panel A, column 2) considers only postgraduate students against the auditor group in order to see the evolution and the differences between the level of students' degree and the auditors. As explained before, postgraduate students present an advanced level of accounting and are more proxy to the audit profession. Moreover, the students from these courses could have experience in audit firms. As previous studies highlight that maturity and accounting education may influence the perception of TFV (see e.g. Rich et al. [18] Rankin et al. [17]), this segmentation allows us to see the evolution between a pre-professional auditor and the auditor.

The dependent variable of the logistic regression model is the 'group of reference' which took the value '1' for auditors and '0' for postgraduate. The division of the students into two groups according to the level of degree adds an interesting result. Looking at the question related to *the integrated concept of TFV in the accounting standards (Q2),* we find statistical differences between auditors and postgraduates ($p < 0.05$). It seems that the level of studies and the maturity of the students have an influence on the perception related to *if the TFV should be abandoned in European accounting due to the fact that it is not important.* The significance of the rest of the variables and the odds ratios are consistent with the estimation of model 1 and the difference of opinion is especially significant between postgraduates and auditors in the case of fines for non-compliance ($p < 0.01$).

Finally, Table 6, Panel B provides the results of the multinomial logistic model. The dependent variable gathered maturity and accounting education and is coded as one, two, or three according to the age of the participant (i.e., between 17–25, between 25–45, and more than 45). The reference group is the youngest group (age between 17–25). The evidence shows that the maturity of participants has an influence on the perception of TFV. Interestingly, there were no statistical differences between the first two groups ($p > 0.05$) but there are statistical differences between the first and the third group. That is, we find that the level of maturity and accounting education of participants affects the questions associated with compliance ($p < 0.10$), definition ($p < 0.10$), and the relationship of TFV and FV ($p < 0.05$) when the participants between 17–25 are compared with the participants older than 45. Furthermore, the differences that we have found in the univariate analysis are driven mainly by the differences between the youngest and the oldest participants.

We can interpret these results as a logical learning curve in the education process of accounting. The key to obtaining a qualitative assessment (such as the interpretation of TFV, a complex concept) depends on their understanding of the concept and the level of maturity in which the concept is studied.

4.3 Sensitivity Analysis

This study hypothesizes that there is a different perception of the meaning of TFV between students and auditors. We also argue that age and accounting education influence the perception of TFV. In this section we report the results from various sensitivity and robustness tests. First, we repeated all tests using the three groups of participants: undergraduates, postgraduates and auditors. Due to limited space, we only show the questions examined in the multivariate analysis.

Briefly, we comment on the main results reported in Table 7. In question Q3 related to the compliance of TFV (whether there is an understanding in Spain to override one or several accounting standards in order to show a TFV, if by using them the TFV objective is not achieved), we observe statistically significant differences depending on the group of participants ($p < 0.05$). We also find statistical differences between the three groups of participants ($p < 0.00$) in question Q7 (opinion on the enforcement of the TFV that in some situations implies penalties, compensations or fines). Finally, we detect statistical differences ($p < 0.00$) in question Q8 relating to whether a detailed definition of the TFV would take away from the efficiency in its application (panel C) and in question Q14 relating to weather the introduction of IFRS (IASB), give the TFV less importance. These results reinforce the evidence found in this paper.

Second, we repeated all the tests using different age and accounting education segmentation. We investigated the effects the difference in the perception of the concept TFV has due to the degree of maturity of each group of participants (differences intra-group). That is, the differences in maturity in the group of students and the group of auditors. Results not reported. Also we use the three age groups used in the previous test and examine the differences driven by professional status. The section between 25–45 years provides special interest because it engages in a high representation from each group of participants (undergraduates, postgraduates and auditors). Table 8 shows only the results of the questions examined in the multivariate analysis due to limited space.

Table 8 shows that the degree of maturity and experience is important in the perception of TFV. For example, in question Q2 related to if the TFV is a foreign concept or should be abandoned in European accounting, we can see how the percentage decreases from undergraduates (12.00%) to auditors (0%) in question Q2. Similar evidence is found in most of the remaining questions. These results reinforce the evidence found in this paper.

Table 7 Test of difference in the perception of TFV between groups

	Undergraduate					Postgraduate				Auditors				Differ. (1–2–3) Kruskal Wallis
		N	%agree	Mean	SD	N	%agree	Mean	SD	N	%agree	Mean	SD	p-value
Integration	Q2	177	8.47	1.62***	1.02	147	6.80	1.58***	0.94	90	1.11	1.40***	0.67	0.14
Compliance	Q3	177	45.76	3.07	1.11	147	49.66	3.12	1.14	90	66.67	3.38 ***	1.24	0.04
	Q7	177	20.90	2.51***	1.17	147	29.25	2.65***	1.24	90	54.44	3.30 **	1.36	0.00
Definition	Q8	177	33.90	2.96	1.08	147	27.21	2.76**	1.12	90	25.56	2.61***	1.16	0.00
TFV&FV	Q14	177	17.51	2.39***	1.13	147	16.33	2.62***	1.06	90	11.11	2.04 ***	1.07	0.00

The Table shows the percentages of agreement (4 or 5 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column shows the *p-value* calculated from the Kruskal–Wallis test

Table 8 Test of difference in the perception of the meaning of the concept TFV in the section between 25–45 years

	Undergraduate					Postgraduate				Auditors				Differ. (1–2–3) Kruskal Wallis
		N	%agree	Mean	SD	N	%agree	Mean	SD	N	%agree	Mean	SD	p-value
Integration	Q2	50	12.00	1.82***	1.22	74	10.81	1.70***	1.07	40	0.00	1.43***	0.64	0.56
Compliance	Q3	50	52.00	3.18	1.06	74	41.89	3.08	1.20	40	62.50	3.30	1.16	0.62
	Q7	50	26.00	2.72*	1.18	74	33.78	2.78	1.32	40	50.00	3.08	1.16	0.32
Definition	Q8	50	52.00	3.40***	0.97	74	25.68	2.72 **	1.20	40	25.00	2.73*	0.96	0.00
TFV&FV	Q14	50	26.00	2.64**	1.16	74	24.32	2.70 **	1.18	40	10.00	2.23 ***	0.92	0.09

The Table shows the percentages of agreement (4 or 5 in the Likert scale), the mean and the standard deviation. The asterisks above the mean figure measure the probability that this average is significantly different from 3, with a level of significance of 1 per cent (***), 5 per cent (**) or 10 per cent (*). The last column shows the *p-value* calculated from the Kruskal–Wallis test

5 Conclusions

This research has been done primarily to evaluate the state of opinion of students and auditors in our study on the TFV as a EU mandatory accounting principle included in local accounting law since 1988, following the accounting Directives. For a codified legal based accounting system it is valuable to know the level of implication of the actual and future accounting agents in reaching the objective of giving a TFV in financial reporting by companies. Besides this main goal, we try to identify if there are differences, due to the effect of experience or the professional responsibilities undertaken, between students (pre and postgraduates) and auditors in public practice.

The identification of differences is important in order to improve the application of the concept in practice and its overall learning process. The EU Directives offer a guide (as shown in graph 1) of how to proceed in applying the concept correctly. The objective would be to minimise the differences between the actors but always in the sense of obtaining the correct use of the concept according to the law. This may mean incorporating changes for auditors or students depending on the questions involved. As explained in the article we include the expected responses from the participants to comply with EU legislation and deviations from these should be eradicated. It is also our intention from the study to consider ways whereby students may be brought nearer to the correct application of the concept during the learning process. An incorrect application of this important accounting concept could have important economic consequences should cases be taken to the courts or on the level of dependence that users place on the accounting information when taking important economic decisions.

One of the most interesting conclusions reached is the high level of acceptance by students and auditors of the objective of the TFV as the main objective for financial reporting. This implies the conformity with the identification of TFV with the principle of substance over form, which constitutes the guidance given in the last accounting reform of the Code of Commerce (in 2007) to reach the TFV in practice. There were no significant differences observed in the responses due to the maturity or accounting education of the participants. Nevertheless, this interesting general result needs to be considered in detail to explain the circumstances surrounding such conformity.

A more important conclusion however is that auditors reflected that the TFV is always obtained by following the accounting standards which would go against the override provision. However in other questions they demonstrated that to show a TFV required more than the accounting standards in vigour at the time which is a contradiction. They also confided that where two or more TFV's were possible the one closest to the accounting standards would be the most correct. This shows that auditors understand the process but clearly they are more comfortable following the standards. This could indicate a practice that is not fully coherent with EU legislation and by applying it in this way could have consequences on the financial information.

The non-professionals give similar answers but not to the same degree but did not give a clear answer to whether the TFV is obtained always by following the

accounting standards. The key conclusion here is that although the override provision for complying with TFV is understood, both non-professionals and auditors would prefer not to use the override provision and comply strictly with the accounting standards. The influence of maturity and professional status of the participants were important here. Once again in this case there would be a need to reinforce the correct application of the TFV.

Another central conclusion is the plea by participants for a definition of TFV. Although all the participants accept the TFV objective as part of the local accounting system, and understand that the attainment of such an objective could imply more than the mere compliance with the accounting standards in force, they appeal for a more detailed definition of the TFV. This could be interpreted as a lack of conformity with the actual situation, because most of the participants realise the difficulty to define TFV but also the response would be more coherent in a codified legal system.

A further fundamental conclusion is the observation that less mature participants support the imposition of penalties or fines for non-compliance with the TFV objective, the more mature participants reject this possibility. This is an interesting conclusion because it shows that auditors are not favourable of imposing a fine when TFV has not been achieved but the accounting standards are complied with. This is in line with auditors and more mature participants being in favour of following only the standards in order to reach the TFV and their need for a definition of the concept. However, if there is no fine for non-compliance then the override condition is useless, it becomes compliance with standards irrelevant of whether the TFV is obtained. This would not be in conformity with the correct application of the concept and it would be interesting to understand why auditors are not in favour of fines for incorrect practice. This is beyond the objectives of this study but could be an interesting focus for future research.

The literature explains that more novice participants tend to be more justice orientated, auditors on the other hand that are dealing with clients have no incentive if their clients can be fined for non-compliance with TFV because it is a reflection on their work and it is easier at the end of the day to follow the accounting standards strictly because it doesn't require any additional judgemental decisions.

Regarding the relation between TFV and the introduction of fair value as an accounting measurement, the higher the professional level the more favourable the reply. This finding is an indirect support to the changes announced, on the dates of the survey, in the General Accounting Plan, adopting the main characteristics of the philosophy of IFRS. It is interesting to see that both, students and auditors, consider that the objective of TFV remains important even after the introduction of fair value in the accounting system.

The study identifies a pattern of change according to the maturity of the participants. The position concerning the TFV objective varies notably with the age of the respondents showing a positive evolution according to age. The more mature the participant the increase in the support that the simple application of accounting standards give the TFV and reject the possibility to establish penalties or fines in the case of non-compliance. On the other hand, the more mature group plead more

strongly for a more detailed definition of TFV and deny the loss of importance of the TFV problem after the introduction of the fair value measurement.

In summary, the two hypotheses formulated in this study cannot be rejected in the area of the integration of the TFV concept in Spanish accounting legislation but the hypotheses can be rejected for most of the questions formulated in the remaining Sections. The effects of professional status and maturity on the perception of TFV are clearly different between the groups of participants in the survey and the level of maturity and accounting education play an important role in the process of perception of this objective for financial reporting. We have no reason to initially believe that this would be different if the study was performed elsewhere because the TFV is a complex accounting concept that needs time to be processed in combination with other learning concepts and through its application in practice. We believe that the perception of the fair presentation override in IASB standards would work in a similar way and would show similar differences between professional status and maturity. We therefore believe that these findings can provide a pathway together with further research in helping to minimise these differences between groups in the perception of this concept and help to achieve a more adequate application of the concept by both actors. This will require correction on the side of both groups of participants in order to eliminate the differences from the required application of the concept according to legislation. This objective must be of special importance firstly to standard setters' but also to professional accounting educators' and to the companies required to show TFV in their accounts'. If the accounting profession wishes to continue with the override clause it must ensure that there is a similar perception on the issue between groups and that this perception is in line with the legal obligation.

Future research in this area would be to examine the evolution of the actors through their professional career. The evolution in the application of TFV as well as other complex concepts after the adoption of IASB accounting rules in Spain could introduce an interesting perspective. Also the relationship could be modelled by variables relating to the accounting standards (specific instructions and requirements, complexity of the accounting standard and amendment issues) and the characteristics of the users. The increasing use of FV opens an area of investigation on whether the changes incorporated in standards continue to provide a TFV of the financial statements. It would also be extremely interesting to analyse participants' ethical intentions with their ethical actions.

Acknowledgements We gratefully acknowledge the help from Alejandro Larriba and several other prestigious researchers for suggestions about the survey and interview design. We also acknowledge the help of several Spanish Universities, the Auditing School of the ICJCE (Instituto de Censores Jurados de Cuentas de España) and AECA (Asociación Española de Contabilidad y Administración de Empresas) for facilitating the distribution of the survey and arranging some interviews. We also appreciate the valuable comments by Marcela Espinosa and the comments from the participants at the 2013 European Accounting Congress (EAA), the 9th Workshop on European Financial Reporting (EUFIN) 2013 and the DGS IV: Decision models in a complex economy 2016.
The authors acknowledge the financial contribution from the Spanish Ministry of Innovation and Science (research projects DER2009-09539, ECO2010-17463, ECO2010-21627, DER2012-33367, ECO2015-66240P, DER2015-67918P), Castilla-La Mancha regional Ministry of Education and Science (research Project POII10-0134-5011) and Alcalá University (research project CCG2014/HUM-036).

References

1. Alexander, D., Jermakowick, E.: A true and fair view of the principles/rules debate. Abacus **42**(2), 132–164 (2006)
2. Alexander, D., Eberhartinger, E.: The true and fair view in the European union. Eur. Account. Rev. **18**(3), 571–594 (2009)
3. Arden, J.: True and fair view: a European perspective. Eur. Account. Rev. **6**(4), 675–679 (1997)
4. Cea, J.L., Vidal, R.: Escenarios de excepción de prevalencia de la imagen fiel sobre los principios y normas contables legales: análisis conceptual y evidencia empírica para las empresas españolas cotizadas en el Ibex 35 [Exceptional scenes of True and Fair View prevalence over legal and accounting principles]. Estudios financieros **308**, 113–150 (2008)
5. Garvey, A.: Los antecedentes de la imagen fiel y su aplicación en España *[The origins of True and Fair View and its application in Spain]*. Dykinson, Madrid (2012)
6. Garvey, A., Gonzalo-Angulo, J.A., Parte, L.: Cognitive Load Theory: Limiting the Gap between Academics and Students in Accounting and Auditing, *Revista de Ciências Empresariais e Jurídicas (RCEJ)*, **28**, 5–28 (2017)
7. Hamilton, G., Ó'Hógartaigh, C.: The third policeman: 'The True and Fair View', language and the habitus of accounting. Criti. Perspect. Account. **20**(8), 910–920 (2009)
8. Houghton, K.: True and fair view: an empirical study of connotative meaning. Account. Organ. Soc. **12**(2), 143–152 (1987)
9. Kirk, N.: Perceptions of the true and fair view concept: an empirical investigation. Abacus **42**(2), 205–235 (2006)
10. Kosmala, K.: True and Fair View or rzetelny i jasny obraz? A survey of polish practitioners. Eur. Account. Rev. **14**(3), 579–602 (2005)
11. Laswad, F.: Perceptions of true and fair view: a New Zealand study. Account. Res. J. **11**(1), (1998)
12. Livne, G., McNichols, M.F.: An empirical investigation of the true and fair override in the United Kingdom. J. Bus. Financ. Account. **36**(1–2), 1–30 (2009)
13. Low, C., Koh, H.C.: Concepts associated with the 'True and Fair View' evidence from Singapore. Account. Bus. Res. **27**(3), 194–204 (1997)
14. Montoya, J., Martínez, F.J., Férnandez-Laviada, A.: La utilización efectiva de los factores cualitativos de la materialidad: un análisis empírico para los auditores de cuentas ejercientes en España. Revista de Contabilidad-Spanish Accounting Review **11**(1), 101–128 (2008)
15. Nobes, C.W., Parker, R.H.: True and fair: a survey of UK financial directors'. J. Bus. Financ. Account. **18**(3), 359–375 (1991)
16. Quick, R., Sánchez, D.: La influencia de las explicaciones de la dirección en la evaluación de los procedimientos analíticos de auditoría. Revista de Contabilidad-Spanish Accounting Review **12**(1), 11–44 (2009)
17. Rankin, M., Silvester, M., Vallely, M., Wyatt, A.: An analysis of the implications of diversity for students' first level accounting performance. Account. Financ. **43**(3), 365–393 (2003)
18. Rich, J.S., Solomon, I., Trotman, K.T.: Multi-auditor judgment/decision making research: a decade later. J. Account. Lit. **16**, 86–126 (1997)
19. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**(2), 257–285 (1988)
20. Van Hulle, K.: The true and fair view override in the European accounting directives. Eur. Account. Rev. **6**(4), 711–720 (1997)

Topics of Disclosure on the Websites: An Empirical Analysis for FinTech Companies

T.-C. Herrador-Alcaide and M. Hernández-Solís

Abstract This paper examines the disclosure of information provided by FinTech companies (FTC) on the website in order to identify the main topics disclosed. Furthermore, the study analyses if the type of service and the geographical area could have some effect on the level of disclosure. The diversity of issues disclosed on the Internet is categorized in items grouped in self-constructed indices in order to identify the structure of the financial and non-financial information reported and furthermore to quantify the volume of disclosure. The empirical study is based on the analysis of the information reported by companies included in two FinTech top-list. Thus, the sample includes 91 businesses from Europe, Asia and North America. The results of the indices indicate that the total volume of information provided on the website is less than the amount reported in other sectors and other type of businesses. The findings also suggest that there are not any effects due to the type of service or the geographical area, effects traditionally associated in the literature with a major or a minor level of disclosure.

Keywords FinTech · FinTech disclosure · Digital financial services · Disclosure indices · FinTech market · FinTech information

1 Introduction

FinTech is a business model which combines Finance and Technology to provide financial solutions [14] by using of a new kind of financial software. Problems of the Small businesses to access to external funding at reasonable rates and existence of a

The authors of this article made similar contributions.

T.-C. Herrador-Alcaide (✉) · M. Hernández-Solís
UNED, Madrid, Spain
e-mail: therrador@cee.uned.es

M. Hernández-Solís
e-mail: montserrath@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_9

strong telecommunications market caused the development of the FinTech business. The mobile phone links different economies in a capital global market and this technological development has boosted innovative concepts of business, such as FinTech. Services through the web are increasing more and more in the banking and financial industry, covering FTC an important quantity of the world financing. For these companies the Internet is the main channel of disclosure.

It is known that FTC are attracting new business ideas and stablishing technology centres such as Silicon Valley or the "FinTechCity" in London. It is estimated that $11 billion were invested in FTC in the global market during the last year in US [15], and in a report about FinTech boom it is shown that these companies raised $8 billion from venture capital in just six months. FTC are considered the new companies for financial and banking services and, for this reason, governments are trying to promote a favourable environment for the FinTech business (COM/2015/063 final).

Since the end of 20th century before this expansion of new financial services companies, the use of the Internet by the companies as usual channel for reporting has been studied. Websites are being used by companies to disclose financial and non-financial information. The idea of greater disclosure is associated with greater transparency and opportunities to take good economic decisions. However, too much information or not structured information could cause a negative effect. Disclose data is not the same as reporting. It is necessary to understand the type and the quantity of information reported on the websites. Thereby, the main problem is the ignorance about the structure of topics disclosed. The first step to homogenize the information for a possible standardization is to know what type of information is disclosed by FTC and whether there are determining factors of it. The academic contribution to resolve this problem has been focused on the analysis of disclosure through experimental studies to quantify the disclosure for different topics [1, 19]. These academic studies have added relevant findings about the optimal set of disclosures to make good business decisions [24], by allowing companies and other stakeholders to know the kind of useful information required. Consequently, the major objective of this article is to examine the information that FTC disclose on the Internet.

Disclosure has been studied for large companies and listed companies, however it can hardly be found studies for small and medium enterprises, generally less regulated. FTC are not generally large listed companies, they usually are start-up companies, and perhaps because of this reason there have not been studies about their disclosure yet. Their customers are small businesses and individual savers who can barely have access to other different information than provided by these companies on their websites. Hence, the information on the website is an important objective for stakeholders and for the society. The previous arguments have lead this paper to improve the theoretical knowledge related to the structure of disclosure provided by FTC on the websites. Beyond that, this paper wants to identify differences in FTC's disclosure by considering geographical areas and type of service in FinTech. Thus, this study contributes to enhance the knowledge about the main topics of disclosure on the website by FTC, the type of information reported in each one, the total volume of information disclosed, and the determining factors for the amount of information disclosed by FTC.

In addition to contributing to fill this knowledge gap, this study points to new lines of research about the adequacy of this information to the demand of the different stakeholders of FTC. This work is important due to the fact that it analyses the issues that the FTC wants to disclose on their website. These issues have not been categorised and thus there is a knowledge gap still unexplored. It should be noted that FTC are developing a highly complex business with a very fast evolvement in recent years and with a great effect on society. In addition, for the nature of the stakeholders, mainly small and medium savers, the improvement of the knowledge about the reporting that these companies make is an important academic and social advance. Furthermore, there are few publications about the FTC, although every day do not stop increasing news about this type of business. Specifically, there are no studies focused on the analysis of the content and classification of the information disclosed by the FTC on the websites. However, this topic has been very studied for other sectors and types of business due to ability to synthesize informative approaches that could be considered by stakeholders in their economic decisions. The structure of this paper is as follows. Section 2 presents the literature review and hypotheses. Section 3 shows the sample. Section 4 provides the research method. Section 5 discusses the results and Sect. 6 presents the conclusions.

2 Theoretical Framework and Hypotheses

FTC provide many type of services in a Peer to Peer manner (P2P) or Business to Business manner (B2B), through financial institutions or directly to customers, but in anyway FTC are providing financial services in an innovative form. The most of FTC were born as start-up companies. The term "Start-up" means a sort of business with a high potential of growing that can be developed in a phased manner. Other feature is that FTC can be born under several legal form. Another important point to define FTC is that workers who normally are entrepreneurs, usually they are also the employees and the owners, at least in an initial phase. Moreover, the digital market has the drawback of the intangible nature of their investment and thus the problems in assessing the risk, arising the FTC as business angels between investors and financial entrepreneurs. A problem underlying the FTC is that the Internet technology still presents many gaps about cybersecurity and this is another feature of FTC. Thus, the main features of FinTech Companies are shown in the Table 1.

Considering previous arguments, this study can be framed within disclosure of information. In this paper different empirical studies on disclosure have been considered which are shown in Table 2.

According the literature, the empirical analysis of the disclosure provided by FTC has been made by using of indices to quantify the amount of information on the website and also to study the relationship of the amount of disclosure with some determining factors of its extension. In this way, the objectives of this study have been focused on the topics of disclosure provided by the FTC considering previous studies for other economic sectors and other type of business, but also the research

Table 1 Characteristics of FinTech companies

By the degree of Innovation	• Entrepreneurship (in a first phase) • Technology-based companies • Start-up company • IT companies (Internet Technology Companies)
By the legal form	• Individual employer • Company form
By the company structure	• More flexible than traditional company • Specialized staff • Founders are entrepreneurs
By the type of services	• Financial and banking services-Digital Technology services • Focus on a few banking services (no more than four) • Easily-explained products
By its relationship with the customers	• B2B versus P2P
By employers/workers	• A few workers (small FT Companies in the first phase) versus many workers (Large FT Companies)

Source Author's development

is based on the main features of FTC discussed above due to that this is a complex model of business. Such thus, the aims of this study are:

1. The identification of the different topics of information provided on the websites by FTC.
2. To summarize these information in indices.
3. The analysis of two hypotheses about of the influence of FinTech service (H1) or FinTech geographical area (H2) on the volume of disclosure, in order to examine the existence of different disclosure cultures depending on each one of these both two factors.

To reach the third objective, the following two hypotheses were tested:

- H1: There is a significant association between the FinTech Service (FinTech subsector) and the amount of voluntary disclosure.
- H2: There is a significant association between the geographic area and the amount of voluntary disclosure.

3 Sample

In this study the companies of two FinTech rankings (top-list) have been examined. The first one is from London and it includes European FTC, the other is from Australia and it includes FTC from several geographical areas. Thus, the sample collects FTC from Europe, the North America and an important part of Asia-Africa. The top-lists

Table 2 Literature about disclosure

Authors	Topics on disclosure
[1] Aksu, and Kosedag, A. (2006)	Association between transparency and disclosure is studied. It is based on the observance of the desirable attributes of the information disclosed by companies listed
[2] Ashbaugh, Johnstone, and Warfield (1999)	The cost and the benefits of IFR (International Financial Reporting). Disclosure before and after of IFR
[4] Bonsón Ponte, and Escobar Rodríguez (2002)	The disclosure on the Internet to goal the association between transparency of disclosure and 3 variables: sector, country and size
[5] Craven, and Marston (1999)	The relationship between Firm's size and disclosure on internet
[6] Debreceny, and Rahman (2005)	It is tested nine hypothesis about the frequency of disclosure (continuous disclosure)
[7] Depoers, F. (2000)	It is studied the effects of several economic determinants on the extent of disclosure in the annual reports for French listed companies, by using an index about financial and non-financial information, finding as an significant factor for extension the size, property and abroad activity
[8] Eng, and Mak (2003)	It is studied the impact of ownership structure and board composition on voluntary disclosure, by using disclosure scores as a proxy to quantify voluntary disclosure, finding a positive association between the prior variables and the voluntary disclosure
[10] Firth (1984)	It is tested whether the volume of disclosure can be linked to the assessment of stock market risk, not finding relevant results between both two variables
[12] Hossain, Tan, and Adams, M. (1994)	Six variables are tested in order to determinate their association with the voluntary disclosure in the annual reports of listed companies. The independent variables tested are firm size, ownership structure, leverage, assets-in-place, size of audit firm, and foreign listing status
[13] Igbal Khadaroo (2005)	It is studied if the expected number of users is an important factor to disclose financial information
[16] Larrán, García-Borbolla, and López, R. (2009)	A temporary disclosure analysis is performed, applying indexes to relate the volume of disclosure to a set of explanatory factors
[19] Nikolaj Bukh, Nielsen, Gormsen, and Mouritsen, J. (2005)	One objective of this article it is to test whether the amount of disclosure could be associated to several factors such as size or age of the company, focusing the study on the information about intellectual capital provided through non-financial reporting
[21] Pérez (2004)	It is studied significant factors for disclosure about competitive advantage of companies through hypothesis on the association between disclosure and size, profitability, company growth, sector, market share and property, by peer
[22] Pirchegger, B., and Wagenhofer, A. (1999)	There is a higher score of free information in large companies than small and medium companies
[25] Urquiza, Navarro, and Trombetta (2009)	Measurement indices of the disclosure are analysed and how the outcome of the measurement of disclosure influences
[26] Xiao et al. (2005)	In this study it was made a survey focus on expert in Internet and accounting variables

Source Author's development

Table 3 Information in the Top-List

	The 50 best FinTech innovators report	The FinTech50 2015
Authors	AWI, KPM AUS and Financial Service Council	Nine Partners. It is made by a team formed by five persons
Objective	The objective is to involve FinTech industry for venture capital invested as a measure of innovation	The objective is to help FinTech companies to raise profile globally, to connect with influencers, investors and funders and generate business
Quantity of companies	50	50
Geographic scope	World wide	Europe
Criteria	4 Criteria: • Total capital raised • Rate of capital raising • Degree of sub industry disruption • Degree of product, service, customer experience and business model innovation	Not specified except for the information about the firm regarding the business area and some economic data about the activity
Information	• Income tax • Revenue from financial services • Percentage that involves financial services	• There are not economic data • There are not template data • A literal description of the type of company. It can infer the type of service, but it is not classified

Source Author's development

are "The FinTech50 2015" and "The 50 Best FinTech Innovators Report". The main characteristics for these rankings are shown in Table 3.

Thus, the sample consists of 100 FinTech companies, 50 for each top-list. Thus, free data provided on the websites are used for the empirical study, such as it is usual in studies about reporting and disclosure on the Internet (See [2, 10, 23]). A total of 100 websites were visited, however, the sample consists of 91 companies because 8 were included in both top-lists and the other one was eliminated for the poor quality of data. These websites were reviewed from July to December 2015.

4 Research Method

This section presents the methodological sequence followed in the empirical study of this research. Thus, it is examined the methodological objectives, the instruments of measurement, and the variables considered: Type of service and geographical area.

4.1 Methodological Objectives

The overall objective of the empirical study of this research is to measure the total volume of information reported by FTC on their website. This general methodological objective has led to the establishment of specific methodological objectives for the empirical work. In this way, the specific objectives have been:

- Disclosure topics identification.
- Establishment of indices to quantify the volume of information.
- Quantification of the information disclosed in each index.
- Testing of the hypotheses that relate to the effect that the FinTech type of service and the geographical area could have extension of the disclosure.

4.2 Disclosure Topics

The top-lists include different types of FinTech business, such as money transfer, platforms for online commerce, platforms for payments through smartphones, granting of loans, as well as other subsectors, loans and payments, and even insurance. All of them have as a common point the financial disintermediation but through the use of high technology on the Internet. FTC of the sample provide several sort of FinTech service (see Table 4). It has been considered previous studies in order to group the different types of FinTech services identified above.

Ashbaugh, Johnstone and Warfield [2] established eleven categories (indices) for a similar research applied to other sector and Debreceny and Rahman [6] only four topics. Regarding Park et al. [20] it is possible to categorize the services in FinTech in four main parts: Banking and Data Analytics, Payments, Capital Markets Tech and Finance Management. The topics have been stablished ad hoc for this study because there are not specific indices in the literature to apply to FTC disclosure, however they were reduced or extended to consider features of FTC. Thus, the information disclosed has been structured around six categories, one for each different topics. Economic data and other information about company have already used in disclosure indices in the most of previous studies [8]. Therefore, this research includes indices about "Economic Data", "Company data", "Staff" and "Partners". In this study two new and specific indices were made, one to measure the information about "Service in FinTech" and the other about the information about "Cybersecurity", both topics are closely linked to FinTech business. These two last indices are two important contributions to improve the knowledge of information disclosed on the websites by FTC.

4.3 Partial Indices and Global Index

The volume of information disclosed is measured by indices. In experimental studies about disclosure many researchers have applied items grouped in indices [3, 9, 17].

Table 4 Service in FinTech

The FinTech50 2015	Thew 50 best FinTech innovators report
Payments	
Financial trading connections	Credit decisions
To use the behavioural biometrics to verify identity	To provide electronic payment
To trade on the stock market	
To lead institutional investors to better manage	To connect to the data sources with small businesses
To offer low-cost loans to customers	To provide loans to small business
Crowdfunding	A peer-to-peer lending and crowdfunding
Management of risk	Financial planning
FinTech consulting	
Banking services	
Insurance services	To link start-ups with angel investors
Big Data finance company	Big Data credit assessment platform
P2P lending	
Video Banking, real time marketing, social channels and mobile banking Cybersecurity	To use technology and algorithms to give advice on a service to clients

Source Author's development on Top-List Web

In accounting research these indices have been made as a proxy about the information disclosed by companies [25]. Also the relationship between information disclosure indices and several factors has been studied [4, 21]. This paper has not limited the disclosure indices just to items related to financial aspect but also other non-financial issues, because information demand is related to the total identity for the FinTech business, and this one is extended a different parts and not only financial aspects. The indices have been constructed *ad hoc* for this research as in other studies (see [17]; and others above).

In the empirical study the items to conform the indices have been applied in order to explain more the information disclosed on the Internet, specifically on each FinTech website as in previous studies [18]. Unweighted indices have been used because it is assumed that all of them are equally important [7, 12]. Therefore, neither the items have been weighted nor the indices. Each item can get only value "1" or "0".

The above arguments have led to the construction of a partial index for each topic of disclosure. Thus, each partial *index* (*Ii*) measures the amount of voluntary information in a specific topic. Thus, each one is determinate for this formula:

$$Ii = \frac{\sum_{j=1}^{n} Items}{\sum_{j=1}^{n} Max\ Items}.$$

Also a Total Index (*TI*) has been calculated. The *TI* has been calculated by the quotient between all six partial indices and the maximum value that it could reach for all. The mathematical expression is:

$$TI = \frac{\sum_{i=1}^{6} Ii}{\sum_{i=1}^{6} Max\, Ii}.$$

The partial indices (*Ii*) summary the level of information disclosed in each category of information on each topic. The total index (*TI*) quantifies the total information disclosed by FTC on the website. Also, the *TI* is applied to test the H1 and the H2.

4.4 The Incidence of Type of Service and Geographical Area

Once the *TI* has been calculated for each FTC, the hypotheses about of type of service and the geographical area have been statistically calculated. Since the first studies on disclosure, researchers have tried to specify if there are variables that can influence the disclosure, such as the size of the company, the sector or subsector, the country, and others. In this research has been analysed two variables as potential factors of a major or minor volume of disclosure: The subsector in FinTech and geographical area.

In previous studies the industrial sector has been considered as an important factor to disclose more or less information, because it is considered that there are different practices for each sector [16]. The argument to study this variable is based on that companies in the same industrial sector or service have similar practices in disclosing information. In this study each FinTech service is a different sector into FinTech Business.

Other studies have traditionally considered that the country is another important factor of the amount of disclosure [5, 11], because in different countries could be different cultures of disclosure. Thus, to analyse the influence of the country, the countries have been grouped in geographical areas or regions. Four areas have been stablished in this research: Europe, North America, World Wide and Asia-Pacific.

In summary, the objective is to check if the quantity of information disclosed on the website depends on the FinTech service for the H1 or the geographical area for the H2. The amount of information is measured by the total index for each company (*TI*), which is the dependent variable, and the independent variables are the type of service and the geographical area.

To evaluate the degree of association between a quantitative variable (*TI* in this study) and a categorical variable[1] (Sectors or geographic areas in this research), the inferential statistical is used to compare the distributions of the quantitative variable

[1] A categorical variable classifies individuals into groups.

in each different group established by the categorical variable. If it has three or more categories for the categorical variable, this comparison of means among three or more groups is analysed through a mathematical model that is the Analysis of Variance (ANOVA). If $p \leq 0.05$,[2] then H0 (the null hypothesis) is rejected and thus a dependent association between variables is accepted (H1 or H2 is accepted). The Kruskall–Wallis test is used as a complementary analyse. This is a non-parametric test to compare more than two groups of ranges (medians) and to determine that the difference is random and it is not statistically significant. If $p \leq 0.05$, then the null hypothesis (H0) is rejected and thus a dependent association between variables is accepted. This means that the quantitative variable behaves differently in the different groups established by the categorical variable.

Thus, in this empirical study the rejection of the null hypothesis would mean that the total volume of information reported by the FTC on the website (*TI*) depends on a different behaviour linked to each topic of FinTech service for the first hypothesis. For the second hypothesis the rejection of the null hypothesis would mean that the differences on the volume of information is caused by the existence of a specific culture of disclosure linked to each geographical areas. This is statistically interpreted as the existence of different disclosure behaviours, either by FinTech subsector of service in FinTech or by geographical area.

5 Findings

Two types of findings can be distinguished as a result of this research, those related to the disclosure topics identified, and those related to disclosure factors. The former show descriptive results about the topics of disclosure identified, while the latter show results of variables which can be determinants of the volume of information disclosed by the different FTC groups.

5.1 Findings Relate to Topics

The results of the Table 5 show that the majority of the disclosure is concentrated around the service information (*I1*) and around general data about the company (*I4*). These rates cover around 45% of the information. Next come the other three partial indices covering around 39% information, such as about of the staff (*I3*), partners (*I5*) and cybersecurity (*I6*). It must be noted that there is only a few of voluntary disclosure on economic data (*I2* takes barely a 12%).

The *TI* shows that the score from the majority of FTC is around 36.79%, some minor than the findings for other type of industry.

[2]For social sciences this level of significance is usually accepted.

Table 5 Indices of disclosure by FTC

	% FIRMS	Mean	Sta. Dvt.	Minimum	Maximum
Service in FinTech (I1)	25.00	0.4601	0.20874	0.00	1.00
Economic data (I2)	12.00	0.1261	0.18920	0.00	0.80
Staff (I3)	10.00	0.3886	0.21726	0.00	0.75
Company (I4)	20.00	0.4438	0.20565	0.00	1.00
Partners (I5)	14.00	0.3949	0.43624	0.00	1.00
Cybersecurity (I6)	19.00	0.3940	0.22000	0.00	1.00
Total Index (TI)	**100.00**	**0.3679**	**0.12670**	**0.00**	**0.80**

Source Author's Development own, SPSS

Table 6 Total Index for each FinTech Service (Subsector)

Service provided	Mean	N	Std. Dev
1 Banking and data analytics	0.3754	22	0.12296
2 Payments	0.3678	22	0.10129
3 Capital markets technology	**0.3765**	21	0.16407
4 Financial managements	0.3545	26	0.11915
Total	0.3679	91	0.12670

Source Author's development, SPSS

Regarding to the items contained in each index (see appendix), for *I1* it should be noted that all companies clearly identify the type of FinTech service provided, but almost none offers online related education. Most FTC also identify the Key Staff or executive team in *I3*, as well as the year of foundation and the scope in *I4*, and data protection and IT support for *I6*. For all these items an information volume of more than 50% is given.

5.2 Findings About the Association Between FinTech Service and TI

Relate to the results about the association between of FinTech service and the quantity of disclosure, it can be seen that there are almost no differences in the quantity of information provided by FinTech Companies on the websites (Table 6). The mean is between 0.35 and 0.37 for each type of service provided for FTC, being the companies which give services in capital markets who disclose more information (21 FTC who provided a mean of 37.65%).

After the first descriptive analysis about the descriptive statistics of the *TI* for each group of FTC classified by categorical variable "FinTech service", the hypotheses

Table 7 Levene and ANOVA -Total Index (Grouping variable: FinTech service)

LEVENE					
	Levene	df1	df2	Sig.	
	1.166	3	88	0.327	
ANOVA					
	Sum of squares	Gl	Mean square	F	Sig.
Between groups	0.008	3	0.003	0.152	**0.928**
Within groups	1.453	88	0.017		
Total	1.461	91			

Source Author's development, SPSS

Table 8 Kruskal–Wallis test (Grouping variable: FinTech service subsector)

	Total Index
Chi-Square	0.465
Gl	3
Asymp. Sig.	**0.927**

Source Author's development, SPSS

have been tested by ANOVA and Kruskal–Wallis. The relationship between each group of FTC by service and the volume of information disclosed is shown in Tables 7 and 8.

The p-value in the ANOVA takes $0.928 > 0.05$ (Sig. Level), then the null hypothesis is accepted and the H1 is rejected. Thus, the means between groups are similar in the *TI*. F-Statistic is not significant, thus the "FinTech Service" is not a determining factor of the total volume of disclosure (*TI*). Therefore, FinTech service has no significant influence on the FTC disclosure. This is also confirmed by the Kruskal–Wallis Test [sig $(0.927) > 0.05$]. These results suggest that FinTech service is not a discriminant factor to disclose a different volume of information. Therefore, it cannot be identified different disclosures among different subsector in FinTech. It is not possible to identify different practices or cultures to provide information on the website depending on each sub-industry of FinTech services.

5.3 Findings About the Association Between Geographical Area of FTC and TI

The descriptive statistics about the association between geographical area and the amount of disclosure is shown in Table 9. It can be observed that the majors percentage of FTC are located in Europe and North America, but the mean is not very different among geographical areas, thus the mean takes values between $0.3386 \leq \text{Mean}\ (TI) \leq 0.4078$, by providing the major volume of information for the FTC located in Asia-Pacific.

Table 9 Total Index by FinTech geographical area

Geographic area	Description for the area	Countries inside	FinTech companies	%	Mean	Standard deviation
1 Europe	FinTech companies located in European Union	Amsterdam UK Sweden Netherlands Germany Finland Czech. Republic Ireland Lithuania	44	48.35	0.3386	0.11047
2 North America	FinTech companies located in North-America	USA Canada	28	30.76	0.4066	0.11482
3 Worldwide	Global companies in FinTech with offices in two or more areas	Several countries	8	8.79	0.3847	0.09575
4 Asia-Pacific	FinTech companies located in Asia-Pacific and one company from Africa	Hong Kong Israel Australia Kenia India China	11	12.08	**0.4078**	0.16761

Source Author's development

Table 10 Levene and ANOVA - Total Index (Grouping area: FinTech geographical area)

LEVENE					
	Levene	df1	df2	Sig.	
	0.831	3	87	0.480	
ANOVA					
	Sum of squares	Gl	Mean square	F	Sig.
Between groups	0.098	3	0.003	2.319	**0.081**
Within groups	1.226	87	0.014		
Total	1.324	90			

Source Author's development, SPSS

The H2 is tested by ANOVA and Kruskal–Wallis Test (see Tables 10 and 11).

The p-value in the ANOVA takes $0.08 > 0.05$ (Sig. level), then the null hypothesis is accepted and thus the H2 is rejected. Thus, geographical area has not a significant effect on the FTC disclosure. The difference is not very high, but it is also refuted by Kruskal–Wallis test (p-value is $0.064 > 0.05$). Therefore, the findings suggest that there is not statistical association among the geographical areas of FTC and the amount of voluntary disclosure on the websites. Thus, it cannot be identified the

Table 11 Kruskal–Wallis test (Grouping variable: geographical area)

	Total Index
Chi-Square	7.274
Gl	3
Asymp. Sig.	**0.064**

Source Author's development

existence of a different culture about voluntary disclosure by geographical areas. However, it must be pointed that the rejection of the H2 is made only by a small difference, thus it would be possible to find another results for other future studies.

6 Conclusions and Discussions

FTC do not seem to have a culture of disclosure yet. The findings show that FTC provide just around 36% of the categorized information in the indices. It cannot identify the existence of a homogeneous structure of disclosure and FTC show a lack of habit in the voluntary disclosure of economic-financial data. This disclosure is mainly directed to the promotion of the business. Thus, the information is focused on services in FinTech and general data about the company. It must be emphasized that FTC hardly offer any economic data on the websites, perhaps because they are not so interested on the social reporting and because society does not require this type of information for this business yet. This plus of information could be considered by FTC more a competitive disadvantage in front of their potential competitors than a positive signal for potential investors. As a reflection, it should say that is a concern the fact that FinTech Companies provide just some economic data, because these companies get a large volume of funds of the small businesses and small investors, in general not overly protected by the legal and economic system.

Furthermore, it is possible to conclude that the total volume of disclosure of FTC is not different in Asia, Europe or North America, reporting around 35% to 40% of the usual items. None of these geographical areas show much more information than the other regions. Neither there is a different level of disclosure considering the type of service provided in opposition of the findings on other studies. Therefore, even considering the own limitations of anywhere statistical analysis, it can be stated that neither the type of service (subsector) nor the area are factors that affect the amount of information disclosed by the FTC on the websites.

These findings could contribute to the improvement of the information provided for the stakeholders of FTC extending the transparency of information. FTC should consider the disclosure of a minimum of information in the indices with lower volume, in order to establish a better competitive position in the digital financial market. If FTC from a geographical area provided a plus of information perhaps they would take a better market position in a long term.

Table 12 Items compliance frequency

Item	Compliance freq.	%
Service provided		
Clearly identifies what type of service**	91	100.00**
Gives a free trial	27	29.35
Education and training online for usage	7	7.61
Economic data		
There is a specific section for economic data	3	3.26
Revenue information	5	5.43
Financing information (rounds and important investors)	28	30.43
Audit information	19	20.65
Listed company information	0	0.00
Staff		
Company size or number of employees	27	29.35
Difference between self-employed and employees	1	1.09
Identifies the Key Staff or executive team**	80	86.96**
Governance (People)	36	39.13
General data about the company		
Year of Foundation**	87	94.57**
Offices across the planet	37	40.22
Offers numbers of customers	12	13.04
Worldwide scope**	66	71.74**
Business policy, SCR and transparency	16	17.39
Benefits (medical insurance, free snacks and drinks, casual dress work environment, competitive salary)	28	30.43
Partners		
Gives some information about the current partner	39	42.39
Offering on its web to became a partner	25	27.17
Offer information about important partners	43	46.74
Cybersecurity		
Explains how data are protected when you are using the FinTech platform**	62	67.39**
Ensures a services 24/7 (always there)	11	11.96
Informatics support information**	59	64.13**
Technological information about platform (programming languages used and others)	13	14.13

Source Author's development

Also it would be interesting to study other determining factors in the disclosure of information from FinTech industry, such as the profile for the different stakeholders in FinTech business. Furthermore, it could be interesting to know how the financial

literacy is linked or not to the FTC's customers and other stakeholders. This is a reflection which could be considered in a future regulation for FTC.

One limitation of this research is that the conclusions from the empirical study are not comparable because barely there are researches focused on the analysis of FTC's data, even less about disclosure on the Internet, so thus this research constitutes an important first step in the knowledge of the FTC, under the approach of social role of these companies in the financial system and their goal of equality to access to the funding sources.

Acknowledgements The authors are grateful for the valuable comments and suggestions made by the anonymous referees in the preparation of this article. The authors also want to thank the suggestions made by several researchers in the design of the empirical study.

Appendix: Statistic Frequency for Each Item

The statistic frequency for each item is shown in Table 12.

References

1. Aksu, M., Kosedag, A.: Transparency and disclosure scores and their determinants in the istanbul stock exchange. Corp. Gov. Int. Rev. **14**(4), 277–296 (2006)
2. Ashbaugh, H., Johnstone, K.M., Warfield, T.D.: Corporate reporting on the internet. Account. Horiz. **13**(3), 241–257 (1999)
3. Barrett, M.E.: The extent of disclosure in annual reports of large companies in seven countries. Int. J. Account. **12**(2), 1–25 (1977)
4. Bonsón Ponte, E., Escobar Rodríguez, T.: A survey on voluntary disclosure on the internet: empirical evidence from 300 European union companies. Int. J. Digit. Account. Res. **2**(1), 27–51 (2002)
5. Craven, B.M., Otsmani, B.: Social and environmental reporting on the internet by leading UK companies. In: 22nd Congress of the European Accounting Association, France (1999)
6. Debreceny, R., Rahman, A.: Firm-specific determinants of continuous corporate disclosures. Int. J. Account. **40**(3), 249–278 (2005)
7. Depoers, F.: A cost benefit study of voluntary disclosure: some empirical evidence from French listed companies. Eur. Account. Rev. **97**(2), 245–263 (2000)
8. Eng, L.L., Mak, Y.T.: Corporate governance and voluntary disclosure. J. Account. Public Policy **22**(4), 325–345 (2003)
9. Firth, M.: A study of the consensus of the perceived importance of disclosure of individual items in corporate annual reports. Int. J. Account. **14**(1), 57–70 (1978)
10. Firth, M.: The extent of voluntary disclosure in corporate annual reports and its association with security risk measures. Appl. Econ. **16**(2), 269–278 (1984)
11. Gray, S.J., Meek, G.K., Roberts, C.B.: International capital market pressures and voluntary annual report disclosures by U.S. and U.K. multinationals. J. Int. Financ. Manag. Account. Spring **6**(1), 43–69 (1995)
12. Hossain, M., Tan, L., Adams, M.: Voluntary Disclosure in an emerging capital market: some empirical evidence from companies listed on the Kuala Lumpur stock exchange. Int. J. Account. **29**, 334–351 (1994)

13. Iqbal Khadaroo, M.: Business reporting on the internet in Malaysia and Singapore: a comparative study. Corp. Commun. Int. J. **10**(1), 58–68 (2005)
14. Khianarong, T., Liebenau, J.: Banking on Innovation-Modernisation of Payment Systems. Springer, Berlin (2009)
15. King, A.: FinTech: throwing down the gauntlet to financial services. Unquote Anal. **21** (2014)
16. Larrán, M., García-Borbolla, A., López, R.: La divulgación de información financiera en la web corporativa de empresas cotizadas. In: VII Workshop of Empirical Research in Financial Accounting and III Research Forum of the SJFA (2009). http://www.viiaccountingworkshop.upct.es/papers/bf5ca7c78d2b4ff8cdd89f8c705616ba.pdf
17. Marston, C.L., Shrives, P.J.: The use of disclosure index in accounting research: a review article. Br. Account. Rev. **23**(3), 195–210 (1991)
18. McNally, G.M., Eng, L.H., Hasseldine, C.R.: Corporate financial reporting in New Zealand: an analysis of user preferences, corporate characteristics and disclosure practices for discretionary information. Account. Bus. Res. **13**(49), 11–20 (1982)
19. Nikolaj Bukh, P., Nielsen, C., Gormsen, P., Mouritsen, J.: Disclosure of information on intellectual capital in Danish IPO prospectuses. Account. Audit. Account. J. **18**(6), 713–732 (2005)
20. Park, J.K., Lee, H.S., Kim, S.J., Park, J.P.: A study on secure authentication system using integrated user authentication service. Indian J. Sci. Technol. **8**, 1–6 (2015)
21. Pérez, G.R.: Factores explicativos de la revelación voluntaria de información sobre fuentes de ventaja competitiva empresarial. Span. J. Financ. Account. **122**, 705–739 (2004)
22. Pirchegger, B., Wagenhofer, A.: Financial information on the Internet: a survey of the homepages of Austrian companies. Eur. Account. Rev. **8**(2), 383–395 (1999)
23. Raffournier, B.: The determinants of voluntary financial disclosure by Swiss listed companies: a reply. Eur. Account. Rev. **6**(3), 493–496 (1997)
24. Schadewitz, H.J., Blevins, D.R.: Voluntary interim disclosures, unexpected earnings, and spreads: international evidence. Int. Adv. Econ. Res. **3**(3), 327–327 (1997)
25. Urquiza, F.B., Navarro, M.C.A., Trombetta, M.: Disclosure index design: does it make a difference? Revista de Contabilidad **12**(2), 253–277 (2009)
26. Xiao, J.Z., Jones, M.J., Lymer, A.: A conceptual framework for investigating the impact of the internet on corporate financial reporting. Int. J. Digit. Account. Res. **5**(10), 131–169 (2005)

On the Thin Boundary of the Fat Attractor

Artur O. Lopes and Elismar R. Oliveira

Abstract For, $0 < \lambda < 1$, consider the transformation $T(x) = dx \pmod 1$ on the circle S^1, a C^1 function $A : S^1 \to \mathbb{R}$, and, the map $F(x, s) = (T(x), \lambda s + A(x))$, $(x, s) \in S^1 \times \mathbb{R}$. We denote $\mathscr{B} = \mathscr{B}_\lambda$ the upper boundary of the attractor (known as fat attractor). We are interested in the regularity of $\mathscr{B}_\lambda$, and, also in what happens in the limit when $\lambda \to 1$. We also address the analysis of the following conjecture which were proposed by R. Bamón, J. Kiwi, J. Rivera-Letelier and R. Urzúa: for any fixed λ, C^1 generically on the potential A, the upper boundary $\mathscr{B}_\lambda$ is formed by a finite number of pieces of smooth unstable manifolds of periodic orbits for F. We show the proof of the conjecture for the class of C^2 potentials $A(x)$ satisfying the twist condition (plus a combinatorial condition). We do not need the generic hypothesis for this result. We present explicit examples. On the other hand, when λ is close to 1 and the potential A is generic a more precise description can be done. In this case the finite number of pieces of C^1 curves on the boundary have some special properties. Having a finite number of pieces on this boundary is an important issue in a problem related to semi-classical limits and micro-support. This was consider in a recent published work by A. Lopes and J. Mohr. Finally, we present the general analysis of the case where A is Lipschitz and its relation with Ergodic Transport.

Keywords Ergodic Optimization · Fat attractor · Maximizing probability · Subaction · Discounted method · Ergodic Transport

1 Introduction

Consider, $0 < \lambda < 1$, the transformation $T(x) = d\,x \pmod 1$, where $d \in \mathbb{N}$, a Lipschitz function $A : S^1 \to \mathbb{R}$, and, the map

A. O. Lopes (✉) · E. R. Oliveira
Instituto de Matemática-UFRGS, Avenida Bento Gonçalves
9500, Porto Alegre-RS, Brazil
e-mail: arturoscar.lopes@gmail.com

E. R. Oliveira
e-mail: oliveira.elismar@gmail.com

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_10

$$F(x,s) = (T(x), \lambda\, s + A(x)),\ \ (x,s) \in S^1 \times \mathbb{R}. \tag{1}$$

Note that $F^n(x,s) = (T^n(x), \lambda^n\, s + [\lambda^{n-1}A(x) + \lambda^{n-2}A(T(x)) + \cdots + \lambda\, A(T^{n-2}(x)) + A(T^{n-1}(x))])$. Here we use sometimes the natural identification of S^1 with the interval $[0, 1)$. We will assume that A is at least Lipschitz. In this case results obtained under such hypothesis could also apply to the shift case (in this setting the potential A is defined in the Bernoulli space), that is, for σ instead of T. For certain results in the paper we assume that $A : S^1 \to \mathbb{R}$ is of class C^1 or sometimes C^2.

The structure of the paper is the following: some results are of general nature and related just to the concept of λ-calibrated subaction (to be defined later). In this case you need just to assume that A is Lipschitz (see Sect. 9).

Other results are related to the dynamics of F and to the conjecture presented in [7]. In this case we assume that the potential A satisfies some differentiability assumptions and also the twist condition (to be defined later). The analysis of the regularity of the boundary of the attractor requires the understanding of λ-calibrated subactions. The twist condition assures some kind of transversality condition as we will see.

Note that for x fixed the transformation $F(x, .)$ is bijective over the fiber over $T(x)$. As an illustration we point out that in the case $d = 2$, given (x, z), with $x \in S^1$ and $z \in \mathbb{R}$, consider x_1 and x_2 the two preimages of x by T. Then, $F\left(x_1, \frac{z-A(x_1)}{\lambda}\right) = (x, z) = F\left(x_2, \frac{z-A(x_2)}{\lambda}\right)$.

It is known that the non-wandering set Ω_λ of $F = F_\lambda$ is a global attractor of the dynamics of F: the forward orbit of every point in $S^1 \times \mathbb{R}$ converges to Ω_λ and F is transitive on Ω_λ. In fact, F is topologically semi-conjugate to a solenoidal map on Ω_λ (see Sect. 2 in [7]).

In Fig. 1 we show the points of the attractor in the case of $T(x) = 2x \pmod 1$, $\lambda = 0.51$ and $A(x) = \sin(2\,\pi\, x)$ (see p. 1013 in [37]). In this case the boundary of the attractor is a finite union of smooth curves.

According to [7] Ω_λ is the set of all (x, s) with a bounded infinite backward orbit (i.e., there exists $C > 0$ and (x_n, s_n), $n \in \mathbb{N}$, such that, $F^n(x_n, s_n) = (x, s)$ and $|s_n| < C$, for all, $n \in \mathbb{N}$).

This transformation F is not bijective. Anyway, the fiber over x goes in the fiber over $T(x)$. If $s_1 < s_2$ is such that (x, s_1) and (x, s_2) are in the attractor, then (x, s) is in the attractor for any $s_1 < s < s_2$ (see Sect. 2.2 in [7]). Note that the iteration of F preserves order on the fiber, that is, given x, if $t > s$, then $\lambda s + A(x) < \lambda t + A(x)$.

We denote $\mathscr{B} = \mathscr{B}_\lambda$ the upper boundary of the attractor. We are interested in the regularity of $\mathscr{B}_\lambda$ and also in what happens with this boundary in the limit when $\lambda \to 1$. The upper boundary is invariant by the action of F. The analysis of the lower boundary is similar to the case of the upper boundary and will not be consider here.

We show in a rigorous form explicit examples where this boundary is the union of a finite number of C^∞ curves where the tangent angles are never zero (see Sect. 7.4). We also present numerical simulations showing pictures of the boundary in several different cases.

The study of the dimension of the boundary of strange attractors is a topic of great relevance in non-linear physics [19, 34]. The papers [1, 2, 15, 20, 21] discuss

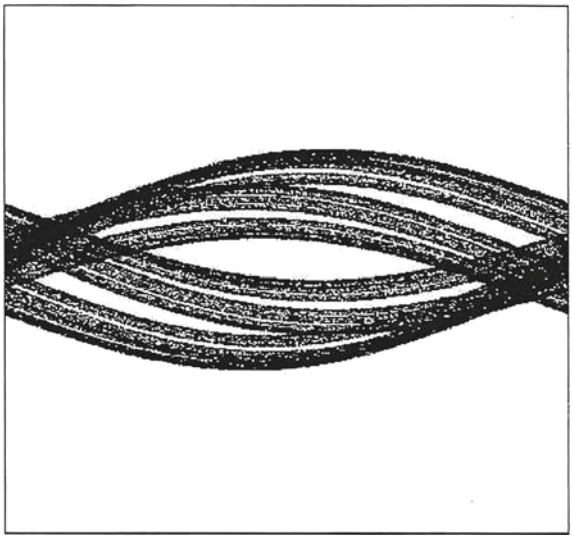


Fig. 1 The fat attractor for the case of $A(x) = \sin(2\pi x)$, $\lambda = 0.51$, and $d = 2$. The picture indicates that the upper boundary is piecewise smooth. It is the envelope of several smooth but non periodic curves

somehow related questions. We want to analyze a case where this boundary may not be a union of piecewise smooth curves.

In Fig. 2 we show de image of the fat attractor for the case of $A(x) = -(x - 0.5)^2$, $\lambda = 0.51$, and $d = 2$. In this case the boundary is the union of two piecewise smooth curves as we will see.

We present in the end of the paper several pictures obtained from computer simulations which illustrate the mathematical results that we present here (see Sect. 8).

We believe that the terminology fat attractor used by M. Tsujii is due to the fact that when $d = 2$, $0.5 < \lambda \leq 0.51$, and $A(x) = \sin(2\pi x)$, then, F is such that there exist a SBR probability which is absolutely continuous with respect to the Lebesgue measure on $S^1 \times \mathbb{R}$ (see Example 1 and Fig. 1 in [37]).

It is known that $\mathscr{B}_\lambda$ is the graph of a Lipschitz function $b_\lambda : S^1 \to \mathbb{R}$ (see [7, 37]) if A is Lipschitz. We will give a proof of this fact later. b_λ will be called the λ-calibrated subaction.

One of our main motivations for the present work is the following conjecture (see [7]): for any fixed λ, generically C^1 on the potential A, the upper boundary $\mathscr{B}_\lambda$ is formed by a finite number of pieces of unstable manifolds of periodic orbits of F.

Recently the paper [29] shows the importance of having a finite number of pieces on this boundary in a problem related to semi-classical limits and micro-support.

We want to also analyze cases where this boundary may not be a union of piecewise smooth curves.

We do not need here the generic hypothesis but we will require the twist condition to be defined later. However, for a generic potential A more things can be said.

The twist hypothesis is natural in problems on Optimization (see [8]) and in problems on Game Theory (see [36]). The twist property for a potential A is presented in Definition 6 in Sect. 3.

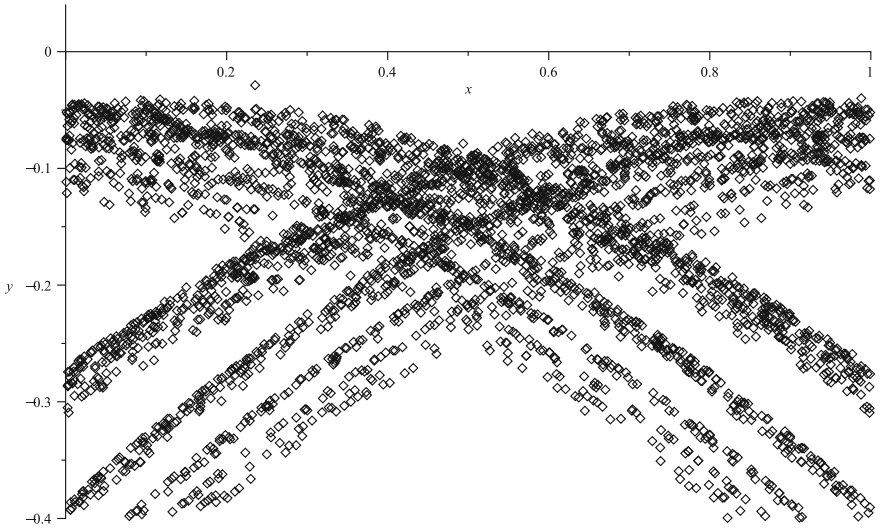


Fig. 2 The fat attractor for the case of $A(x) = -(x-0.5)^2$, $\lambda = 0.51$, and $d = 2$. The picture indicates that the upper boundary is piecewise smooth. It is the envelop of two smooth but non periodic curves

In the same spirit of [30] the idea here is to use an auxiliary family of functions $W_w(x) = W(x, w)$ indexed by $w \in \{1, 2, \ldots, d\}^{\mathbb{N}}$ such that for each w we have $W_w : (0, 1) \to \mathbb{R}$ is, at least C^1 (it is C^2 in the case we consider). This function W of the variable (x, w) is called involution kernel. W_w is not necessarily periodic on S^1 (see pictures on Sect. 8 where a certain S replaces the above W). A natural strategy would be to assume that A satisfies a twist condition and to show that there exists a finite number of points w_j, $j = 1, 2, \ldots, k$, and a corresponding set of real values $\alpha_1, \alpha_2, \ldots, \alpha_k$, such that, for each $x \in S^1$ we have

$$b_\lambda(x) = \max_{j=1,2,\ldots,k} \{\alpha_j + W(x, w_j)\}, \tag{2}$$

where the graph of b_λ is $\mathscr{B}_\lambda$.

In [31] the results assume, among other things, that an special point (the turning point) was eventually periodic. Here we will just use the fact that A satisfies a twist condition. We will show that the conjecture is true when A satisfies a twist condition (see Corollary 2 and comments after Corollary 2 on Sect. 4). We point out that the twist condition is an open property in the C^2 topology. The main problem we analyze here could also be expressed in the C^2 topology.

Expressions of the kind (2) appear in Ergodic Transport (see [28, 30, 31, 33]). Equation (11) just after Theorem 6 describes relation (2) under certain general hypothesis: the Lipschitz setting (see Sect. 9).

We apply all the previous results to the case when A is quadratic in Sect. 7. The main problem we analyze here could also be expressed in the C^2 topology.

In Sect. 9 we describe some general properties related to Ergodic Transport for the setting we consider here.

2 λ-Calibrated Subactions and λ-Maximizing Probabilities

Definition 1 Given a continuous function $A : S^1 \to \mathbb{R}$ and $\lambda \in (0, 1)$, we say that a continuous function b_λ **is a λ-calibrated subaction for** A, if for all $x \in S^1$, $b(x) = \max_{T(y)=x} \{\lambda\, b(y) + A(y)\}$.

A similar concept can be consider when the dynamics is defined by the shift and not T (see Sect. 9).

When A is Lipschitz for each $\lambda \in (0, 1)$ the function b_λ above exist, is Lipschitz and it is unique (see [10, 32]). The existence of such b_λ when A is Lipschitz is also presented in the survey paper [27].

About the interest in such family b_λ we can say that in Aubry-Mather theory and also in Optimization a similar kind of problem is considered in problems related to the so called infinite horizon discounted Hamilton-Jacobi equation. It provides an alternative method for showing the existence of viscosity solutions (see [17, 18, 22]). Thanks to the formal association with Optimization and Economics the analysis of such family b_λ, which takes in account values $\lambda \in (0, 1)$, can be called the discounted problem for the potential A. If A is Lipschitz it is known (see [7]) that the upper boundary of the attractor is the graph of the Lipschitz λ-calibrated subaction $b_\lambda : S^1 \to \mathbb{R}$.

The above result means that if the point $(x, b_\lambda(x))$ is in the upper boundary of the attractor, then, there is a point y such that $T(y) = x$, and $F(y, b_\lambda(y)) = (x, b_\lambda(x))$. In this way the analysis of the dynamics of F on the boundary of the attractor is quite related to the understanding of λ-calibrated subactions.

Note that if b is the λ-calibrated subaction for A, then, $b + \frac{g}{\lambda}$ is the λ-calibrated subaction for $A + \frac{g\circ T}{\lambda} - g$. In order to obtain our main result on the boundary of the attractor we have to investigate properties of λ-calibrated subactions. The three keys elements on our reasoning are: probabilities with support in periodic orbits (see Sect. 2), a relation of the kind (2) for the function b whose graph is the boundary of the attractor (see Sect. 3) and the twist condition (see Sect. 4).

We will present now some general results on λ-calibrated subactions. We denote by $\tau_i, i = 1, 2, \ldots, d$, the inverse branches of T. For each i the transformation τ_i has domain $[i - 1/d, i/d]$. Given x, if i_0 is such that $b(x) = \lambda b(\tau_{i_0}(x)) + A(\tau_{i_0}(x))$, we say $\tau_{i_0}(x)$ realizes $b(x)$ (or, realizes x). We can also say that i_0 is a symbol which realizes $b(x)$. One can show that for $d = 2$, for any Holder A, there exist x such that $b(x)$ has two different $\tau_{i_0}(x)$ realizers. In this way realizers are not always unique. For fixed $x_0 \in S^1$ consider x_1 such that $b(x_0) = \lambda\, b(x_1) + A(x_1)$,

and $T(x_1) = x_0$. Then, there exist a realizer i_0 such that $\tau_{i_0}(x_0) = x_1$. Now take x_2 such that $b(x_1) = \lambda\, b(x_2) + A(x_2)$ and $T(x_2) = x_1$. In the same way as before, there exist i_1 such that $\tau_{i_1}(x_1) = x_2$. In this way get by induction a sequence $x_k \in S^1$ such that $T(x_k) = x_{k-1}$. This also defines an element $a = a(x_0) = (i_0, i_1, \ldots, i_n, \ldots) \in \Sigma = \{1, \ldots, d\}^{\mathbb{N}}$, where $\tau_{i_k}(x_k) = x_{k+1}$. This a depends of the choice of realizers we choose in the sequence of preimages. We say $(x_0, a(x_0)) \in S^1 \times \{1, 2, \ldots, d\}^{\mathbb{N}}$ is an optimal pair. Note that for each $x_0 \in S^1$ there exist at least one optimal pair. For each x_0 we consider a fixed choice $a(x_0)$, and, the corresponding sequence $x_k \in S^1$, $k \in \mathbb{N}$.

Consider the probability $\mu_n = \sum_{j=0}^{n-1} \frac{1}{n}\, \delta_{x_n}$ and μ_λ any weak limit of a convergent subsequence μ_{n_k}, $k \to \infty$. The probability μ_λ on S^1 is T invariant and satisfies

$$\int (b(T(x)) - \lambda\, b(x) - A(x))\, d\mu_\lambda = 0.$$

Note that $b(T(z)) - \lambda b(z) - A(z) \geq 0$ for all $z \in S^1$. In this way **for z in the support of μ_λ** we get the **λ-cohomological** equation

$$b(T(z)) - \lambda b(z) - A(z) = 0. \tag{3}$$

Therefore, μ_λ is maximizing for the potential $A(z) - b(T(z)) + \lambda b(z)$. For z in the support of μ_λ we have that $F(z, b(z)) = (T(z), b(T(z))$. Moreover, in this case

$$b(T(z)) = \max_{T(y)=T(z)} \{\lambda\, b(y) + A(y)\} = \lambda\, b(z) + A(z). \tag{4}$$

Definition 2 Any probability which maximizes $A(z) - b(T(z)) + \lambda b(z)$ among T-invariant probabilities, where b is the λ-calibrated subaction, will be called a λ-maximizing probability for A.

Any μ_λ obtained from a point x_0 and a family of realizers described by a certain $a = a(x_0)$ as above is a λ-maximizing probability for A. Note that μ_λ is not maximizing for A but for the potential $A(z) - b(T(z)) + \lambda b(z)$. General references for maximizing probabilities are [6, 9, 12, 16, 24, 27]. As we are maximizing among invariant probabilities we can also say that μ_λ is maximizing for the potential

$$A(z) + (\lambda - 1)b(z) = A(z) - b(T(z)) + \lambda b(z) + [b(T(z)) - b(z)].$$

Proposition 1 *If z is a point in a periodic orbit of period k and moreover z is in the support of the maximizing probability μ_λ, then the realizer a can be taken as a periodic orbit of period k for the shift σ acting in the Bernoulli space. We call such probability invariant for the shift of dual periodic probability.*

Proof In order to simplify the reasoning suppose $k = 2$. Note that $T(z)$ is also in the support of the maximizing probability μ_λ. In this case, from (3) we have that $b(\,T(\,T(z)\,)\,) = \lambda\, b(T(z)) + A(T(z))$ because $T(z)$ is in the support of μ_λ. From Eq. (4)

$$b(z) = b(\,T(\,T(z)\,)\,) = \max_{T(y)=T^2(z)=z} \{\lambda\, b(y) + A(y)\} = \lambda\, b(T(z)) + A(T(z)).$$

Therefore, $\max_{T(y)=z} \{\lambda\, b(y) + A(y)\} = \lambda\, b(T(z)) + A(T(z))$. In this way we can take the corresponding inverse branch, say a_1, and then, say a_2, and we repeat all the way this choice again and again in order to define $a = (a_1, a_2, a_2, a_4, \ldots) = (a_1, a_2, a_1, a_2, \ldots)$. In this case a is an orbit of period two for σ and the claim is true. In the case $k = 3$, note that if $T^3(z) = z$, we have that $b(\,T^2(\,T(z)\,)\,) = \lambda\, b(T^2(z)) + A(T^2(z))$ and $b(\,T(\,T(z)\,)\,) = \lambda\, b(T(z)) + A(T(z))$, because $T(z)$ and $T^2(z)$ are in the support of μ_λ. In this way we follow a similar reasoning as before and we get an a which is an orbit of period 3 for the shift. Same thing for a periodic orbit with a general k.

As an example of the above, suppose $k = 2$, then there are two periodic orbits of period 3. Take one of them, let us say $\{z_1, z_2, z_3\}$. Suppose $T(z_1) = z_2$, $T(z_2) = z_3$ and $T(z_3) = z_1$. Given z_1 there exists a_1 such that $\tau_{a_1}(z_1) = z_3$. Given z_3 there exists a_2 such that $\tau_{a_2}(z_3) = z_2$. Finally, given z_2 there exists a_3 such that $\tau_{a_3}(z_2) = z_1$. Then, in this case, $a = (a_1, a_2, a_3, a_1, a_2, a_3, a_1, \ldots)$ is in the support of the dual periodic probability for $\{z_1, z_2, z_3\}$. The set $\{a, \sigma(a), \sigma^2(a)\}$ defines a periodic orbit of period 3 for the shift σ.

Definition 3 We denote by R the function $R = A - b \circ T + \lambda b \geq 0$ and call it the rate function.

For fixed λ, by the fiber contraction theorem [35] (Sect. 5.12 p. 202 and Sect. 11.1 p. 433) we get that the λ-calibrated subaction $b = b_{\lambda,A} = b_A$ is a continuous function of A in the C^0 topology. Moreover, the function $b = b_{\lambda,A}$ is a continuous function of A and λ.

Taking $\lambda \to 1$ will see now that we will get results which are useful for classical Ergodic Optimization.

We denote $m(A) = \sup\{\int A\, d\,\nu$, among T-invariant probabilities $\nu\}$.

We call maximizing probability for A any ρ which attains the supremum $m(A)$. We denote any of these ρ by μ_A. A continuous function $U : S^1 \to \mathbb{R}$ is called a calibrated subaction for $A : S^1 \to \mathbb{R}$, if, for any $y \in S^1$, we have

$$U(y) = \max_{T(x)=y} [A(x) + U(x) - m(A)]. \tag{5}$$

If $b_\lambda^{\&} = b_\lambda - \max\, b_\lambda$, then, of course, μ_λ is maximizing for the rate function potential $A(x) - b_\lambda^{\&}(T(z)) + \lambda b_\lambda^{\&}(z)$. It is known (see [5, 10, 27, 32]) that b_λ is equicontinuous in λ, and, any convergent subsequence, $\lambda_n \to 1$, satisfies $b_{\lambda_n}^{\&} \to U$, where U is a calibrated subaction for A.

Assuming the maximizing probability for A **is unique** (a generic property according to [12]), it is known (see [16] Sect. 4), that when $\lambda \to 1$, we get that $b_\lambda^{\&} \to U$, where U is a (the) calibrated subaction for A. In this way we can say that the family $b_\lambda^{\&} \to U$ selects the calibrated subaction U via the discounted method. Assuming that the maximizing probability μ_A is unique, when $\lambda \to 1$, we get that $\mathscr{B}_\lambda$ (after the subtraction of $\max b_\lambda$) converges to the graph of the calibrated subaction for A in the C^0-topology (Fig. 3).

Even if the maximizing probability for A is not unique there exist anyway a unique special limit subaction when $\lambda \to 1$ (see [23]). That is, there exist a selection on the discounted method for any Holder potential A (the potential do not have to be generic). In other words, given the potential A, the limit of any sequence $b_{\lambda_n}^{\&}$, $n \to \infty$, $\lambda_n \to 1$, will be a unique special subaction U for A (independent of the sequence). is boundary $\mathscr{B}_\lambda$, when $\lambda \to 1$.

Note that in any case it is true the relation: for any z

$$b^{\&}(T(z)) - \lambda b^{\&}(z) + (1-\lambda)(\max b_\lambda) - A(z) \geq 0.$$

We point out that in classical Ergodic Optimization, given a Lipschitz potential $A : S^1 \to \mathbb{R}$, in order to obtain examples where one can determine explicitly the maximizing probability or a calibrated subaction, it is necessary to know the exact value the maximal value $m(A)$ (see Eq. (5)). In the general case this is not an easy task and therefore any method of approximation of this maximal value or associated subaction is helpful. The discounted method provides approximations b_λ, $\lambda \in (0, 1)$, in the C^0-topology, of calibrated subactions for A via the Banach fixed point theorem, that is, via a contraction in the set of continuous functions in the C^0-topology. You take any function, iterate several times by the contraction and you will get a function $\hat{b}_\lambda$ which is very close to the λ-calibrated subaction. If λ is close to 1, then, the corresponding $\hat{b}_\lambda^{\&}$ is close to a classical calibrated subaction U. In all this is not necessary to know the value $m(A)$. However, when λ becomes close to the value 1 this contraction becomes weaker an weaker.

Theorem 1 claims that for a generic potential A the maximizing measure for A is attained by a λ-maximizing probability for λ in an interval of the form $[1-\varepsilon, 1]$. Thanks to all that one can apply our reasoning for a λ is fixed and close to 1. In the discounted method taking $\lambda \sim 1$ the procedure also provides a way to approximate the value $m(A)$ as we will see later.

It is known (see [27]) that $(1-\lambda)(\max b_\lambda) \to m(A)$. Then, as we said before, one can get an approximate value of $m(A)$ by the discounted method.

It is easy to see that close by the periodic points the graph of b is a piece of unstable manifold for F (see Fig. 4). We point out that for a generic Lipschitz potential A the maximizing probability for A is a unique periodic orbit (see [11]).

Now we state a result which is new in the literature.

Theorem 1 *If ν is a weak limit of a converging subsequence $\mu_{\lambda_n} \to \nu$, $\lambda_n \to 1$, then, ν is a maximizing probability for A. For a generic Lipschitz potential A there exist an ε, such that for all $\lambda \in (1-\varepsilon, 1]$, the λ-maximizing probability has support*

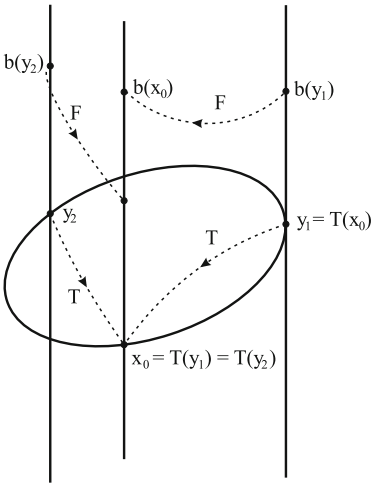


Fig. 3 A geometric picture of the λ-calibrated property of b. The graph of b describes the upper boundary of the attractor

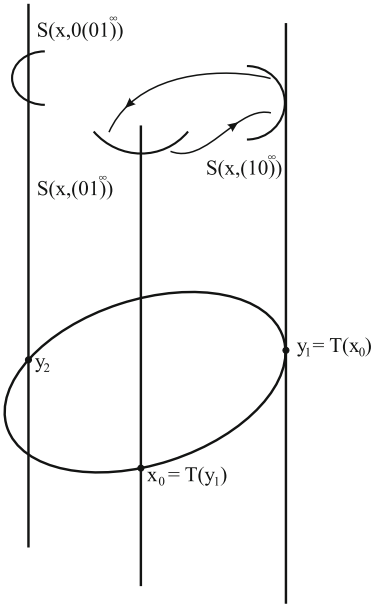


Fig. 4 The unstable manifolds of a point of period two for F

in the periodic orbit which defines the maximizing probability for A. If the potential A is of class C^1 the same is true on the C^1 topology.

Proof Consider a subsequence $\mu_{\lambda_n} \to \nu$, $\lambda_n \to 1$. Such ν is clearly invariant. Suppose by contradiction that for some $\varepsilon > 0$ there exists an invariant μ such that $\int (A - U \circ T + U)\, d\nu + \varepsilon < \int (A - U \circ T + U) d\mu$, then, for any n large enough

we have $\int (A - b_{\lambda_n} \circ T + \lambda_n b_{\lambda_n})\, d\mu_{\lambda_n} + \varepsilon/2 < \int (A - b_{\lambda_n} \circ T + \lambda_n b_{\lambda_n}) d\mu$, and, we reach a contradiction.

Now, for a generic potential it is known that the maximizing probability for A is a unique periodic orbit (see [11]). Therefore, $\mu_\lambda \to \nu$, when $\lambda \to 1$. From the continuous varying support (see [12]) if $\mu_\lambda \to \nu$ and ν is periodic orbit, then, there exist an $\varepsilon > 0$ such that for $\lambda \in (1-\varepsilon, 1]$ the probability $\mu_\lambda = \nu$. If the potential A is of class C^1 one can do the following: since μ_λ is maximizing for $A + (1-\lambda b)$, which is Lipschitz-close to A, then, when λ is close to the value 1, we apply the continuous varying support property in order to get a Lipschitz subaction and perturb in the same way as in [11] in order to get an approximation by another Lipschitz potential with support in a unique periodic orbit. This potential can be approximated once more in the C^1 topology and again in his way by the continuous varying support we get C^1 potential with support in a periodic orbit. Then, the same formalism as above can be applied.

If $\mu_\lambda \to \mu_A$, when $\lambda \to 1$, we say that μ_λ selects the maximizing probability μ_A. In the present case for a generic A there is a selection via the discounted method.

Remark 1 We point out the final conclusion: for a generic A we have that for λ close to 1 the maximizing probability μ_λ is a periodic orbit. Moreover, by Proposition 1 the realizer a for a point x in the support of μ_λ can be taken as a periodic orbit (with the same period) for the shift σ.

An interesting example is $A(x) = -(x-0.5)^2$ and $T(x) = 2\,x \pmod 1$, which has a unique maximizing probability μ_A which is the one with support in the periodic orbit of period 2 according to Corollary 1.11 in [25] (see also [26]). Therefore, the corresponding λ-maximizing probability μ_λ converges to this one. In fact, there is an ε such that if $1 - \varepsilon < \lambda < 1$, then μ_λ has support in this periodic orbit. This example will be carefully analyzed in the lasts sections of the paper.

3 The λ-Calibrated Subaction as an Envelope

Consider (as M. Tsujii in expression (3) p. 1014 [37]) the function $S : (S^1, \Sigma) \to \mathbb{R}$, where $\Sigma = \{1, 2, \ldots, d\}^{\mathbb{N}}$, given by

$$S_{\lambda,A}(x, a) = S(x, a) = \sum_{k=0} \lambda^k A(\tau_{k,a} x), \tag{6}$$

where $(\tau_{a_{k-1}} \circ \cdots \circ \tau_{a_0})(x) = \tau_{k,a} x$ and $a = (a_0, a_1, a_2, \ldots)$. For a fixed a the function $S_{\lambda,A}(., a)$ is continuous up to the point 0 (see several computer simulations in Sect. 8 and the explicit expression for the quadratic case in Sect. 7.4). Note that if A is of class C^2, then for a fixed a the function $S(., a)$ is smooth up to the point 0 in S^1, if $1 > \lambda > \frac{1}{d}$ (see in p. 1014 the claim between expressions (3) and (4) in [37]).

We point out that the upper boundary of the attractor is periodic but each individual $S(x, a)$ as a function of x is not (see worked examples in the end of the paper). Note also that for λ and a fixed the function $S_{\lambda,A}(, a)$ is linear in A. All x has a corresponding $a = a(x)$ such that $b(x) = S(x, a)$. Indeed, for the given x take i_0 such that $b(x) = \lambda b(\tau_{i_0}(x)) + A(\tau_{i_0}(x))$, then, take i_1 such that $b(\tau_{i_0}(x)) = \lambda b((\tau_{i_1} \circ \tau_{i_0})(x)) + A((\tau_{i_1} \circ \tau_{i_0}(x))$, and so on. In this way we get $a = (i_0, i_1, i_2, \ldots)$. This a is not necessarily unique. We call any such possible $a(x)$ a **realizer for** x. Note that

$$b(x) = \lambda\,[\,\lambda u((\tau_{i_1} \circ \tau_{i_0})(x)) + A((\tau_{i_1} \circ \tau_{i_0}(x))\,] + A(\tau_{i_0}(x)) =$$

$$\lambda^2 u((\tau_{i_1} \circ \tau_{i_0})(x)) + \lambda\, A((\tau_{i_1} \circ \tau_{i_0}(x)) + A(\tau_{i_0}(x)) =$$

$$\lambda^n u((\tau_{i_{n-1}} \circ \cdots \circ \tau_{i_1} \circ \tau_{i_0})(x)) +$$

$$\lambda^{n-1}\, A((\tau_{i_{n-1}} \circ \cdots \circ \tau_{i_1} \circ \tau_{i_0})(x)) + \cdots + \lambda\, A((\tau_{i_1} \circ \tau_{i_0}(x)) + A(\tau_{i_0}(x)).$$

Taking the limit when $n \to \infty$ we get $b(x) = S(x, a)$.

From the construction we claim that for any other $c \in \{1, 2, \ldots, d\}^{\mathbb{N}}$ we have $b(x) \geq S(x, c)$. Indeed, consider $z(x) = \limsup_{n\in\mathbb{N}}\{\lambda^{n-1} A((\tau_{i_{n-1}} \circ \cdots \circ \tau_{i_1} \circ \tau_{i_0})(x)) + \cdots \quad \cdots + \lambda\, A((\tau_{i_1} \circ \tau_{i_0}(x)) + A(\tau_{i_0}(x)) \mid (i_0, i_1, \ldots, i_{n-1}) \in \{1, 2, \ldots, d\}^n\}$, and, the operator $\mathscr{L}_\lambda(v)(x) = \sup_{i=1,2\ldots,d} [A(\tau_i(x)) + \lambda v(\tau_i(x))]$. Then, $\mathscr{L}_\lambda(z) = z$. It is known that b is a fixed point for $\mathscr{L}_\lambda$ (see Sect. 3 in [27], or Sect. 2 in [5]). From the uniqueness of the fixed point it follows the claim. Therefore, we get from above the following result which we call the Envelope Theorem.

Theorem 2 *$b(x) = b_{\lambda,A}(x) = \sup_{c\in\{1,2,\ldots,d\}^{\mathbb{N}}} S(x, c) = S(x, a(x))$, where $a(x)$ is a realizer for x.*

As the supremum of convex functions is convex we get that for λ fixed the function $b_{\lambda,A}$ varies in convex way with A.

Figure 2 suggests that $b(x)$ is obtained as $\sup\{S(x, w_1), S(x, w_2)\}$, where, w_1, w_2, is in Σ. Later we will show that in several interesting examples w_1, w_2 are in a periodic orbit of period 2 for the shift σ. As we said before in the introduction we point out again here (in a more precise way) that Fig. 1 in [37] suggests that $b(x)$ is obtained as $\sup\{S(x, w_i), i = 1, 2, 3, 4\}$, where, w_i, $i = 1, 2, 3, 4$, are in Σ. Note that the potential A in that case is conjectured to have a maximizing probability in an orbit of period 4 (see [14]).

Note that if A is of class C^2, then, $S_a : (0, 1) \to \mathbb{R}$ is of class C^2. We define $\pi(x) = i$, if x is in the image of $\tau_i(S^1 - \{0, 1\})$, $i \in \{1, 2, \ldots, d\}$.

Note also (see (7) p. 1014 in [37]) that $S(T(x), \pi(x)a) = A(x) + \lambda\, S(x, a)$. Or, in another way, for any $a = (a_0, a_1, \ldots)$ we have that $S(x, a) = A(\tau_{a_0}(x)) + \lambda S(\tau_{a_0}(x), \sigma(a))$. This also means that $\phi(x, a) = (x, S(x, a))$ is a change of coordinates from F to $\mathbb{T}(x, a) = (T(x), \pi(x)a)$ [37]. $\mathbb{T}(x, a)$ is forward invariant in the upper boundary $\mathscr{B}$ of the attractor. Note also that $\mathbb{T}^{-1}(x, a) = (\tau_{a_0}(x), \sigma(a))$ (when defined).

Definition 4 Consider a fixed $\overline{x} \in S^1$ and variable $x \in S^1$, $a \in \{1,2\}^{\mathbb{N}}$, then we define

$$W(x,a) = S(x,a) - S(\overline{x},a). \tag{7}$$

We call such W the λ-involution kernel for A.

Note that for a fixed $W(x,a)$ is smooth on $x \in (0,1)$. From the above definition we get,

$$\lambda W(\tau_{a_0}(x), \sigma(a)) - W(x,a) =$$

$$[\lambda S(\tau_{a_0}(x), \sigma(a)) - \lambda S(\overline{x}, \sigma(a))] - [S(x,a) - S(\overline{x},a)] =$$

$$[\lambda S(\tau_{a_0}(x), \sigma(a)) - S(x,a)] - [\lambda S(\overline{x}, \sigma(a)) - S(\overline{x},a)] =$$

$$-A(\tau_{a_0}(x)) + [\lambda S(\overline{x}, \sigma(a)) - S(\overline{x},a)].$$

Note that $[\lambda S(\overline{x}, \sigma(a)) - S(\overline{x},a)]$ just depend on a (not on x).

Definition 5 If we denote $A^*(a) = [\lambda S(\overline{x}, \sigma(a)) - S(\overline{x},a)]$, we get the **$\lambda$-coboundary equation**: for any (x,a)

$$A^*(a) = A(\tau_{a_0}(x)) + [\lambda W(\tau_{a_0}(x), \sigma(a)) - W(x,a)].$$

We say that A^* is the λ-dual potential of A.

Note that A is defined for the variable $x \in S^1$ and A^* is defined for a which is in the dual space Σ. The above definition is similar to the one presented in [4, 13, 31, 32]. Note that we have an explicit expression for the λ-involution kernel W which appears in the above definition.

Below we consider the lexicographic order in $\{1,\dots,d\}^{\mathbb{N}}$.

Definition 6 We say that A satisfies the twist condition, if an (then, any) associated involution kernel W, satisfies the property: for any $a < b$, we have

$$\frac{\partial W}{\partial x}(x,a) - \frac{\partial W}{\partial x}(x,b) > 0.$$

Note that this condition does not imply that there is a uniform positive lower bound for $\frac{\partial W}{\partial x}(x,a) - \frac{\partial W}{\partial x}(x,b)$ when $b > a$.

It is equivalent to state the above relation for S or for W. An important issue is described by Proposition 2.1 in [13] which basically says that (in our context) in the case W satisfies the twist property, then association x to a realizer $a(x)$ is monotone, where we use the lexicographic order in $\{1,\dots,d\}^{\mathbb{N}}$. See also Proposition 8 in the last section. This is not exactly, but very close, of saying that the support of the optimal plan probability for W is a graph. In [31] the question about the property of cyclically monotonicity (in the support) is addressed.

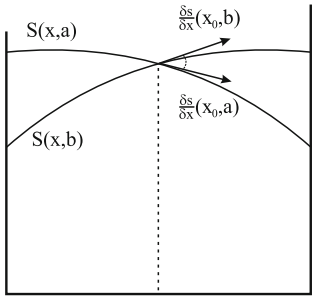


Fig. 5 Under the twist condition the way the two graphs cut is compatible with the inequality $b < a$ (see Proposition 2)

4 Geometry, Combinatorics of the Graphs of $S(., a)$ and the Twist Condition

Remember that $S(x,a) = \sum_{k=0} \lambda^k A(\tau_{k,a} x)$, and $W(x,a) = S(x,a) - S(\overline{x},a)$. Moreover, one can get the calibrated subaction via the **superior envelope** $b(x) = b_{\lambda,A}(x) = \sup_{a\in\Sigma} S(x,a) = S(x,a(x))$. In this case $a(x) \in \Sigma$ is called an optimal symbolic element for x (possibly not unique). Under the twist condition, two graphs cut one each other in a compatible way with the inequality $a < b$. The envelope result, assures that, if the family $S(x,\cdot)$ is continuous in Σ, then $\frac{\partial b_{\lambda,A}}{\partial x}(x,a) = \frac{\partial S}{\partial x}(x,a)$, for every optimal a. Thus, if A is twist the optimal symbolic element is unique in every differentiable point of $b(x) = b_{\lambda,A}(x)$. **The two graphs on the Fig. 5 can not cut twice by the twist property. This is the purpose of the next results. Note, however, that the graph of one $S(., c)$ will be intersected by an infinite number of other graphs of $S(., d)$.**

We will study now some additional properties of the family of maps $S(x,a)$. The first step is to consider some especial functions.

Definition 7 For a fixed pair $a, b \in \Sigma$ we define $\Delta : S^1 \times \Sigma \times \Sigma \to \mathbb{R}$ by $\Delta(x,a,b) = S(x,a) - S(x,b)$, that is C^2 smooth on $x \in (0,1)$.

Computing this derivatives we get $\Delta'(x,a,b) = S'(x,a) - S'(x,b)$ and $\Delta''(x,a,b) = S''(x,a) - S''(x,b)$, thus we get two consequences. The first: if A is twist and $a \neq b$ then $\Delta'(x,a,b) \neq 0$, more precisely, if $a < b$ then $\Delta'(x,a,b) > 0$ else, if $a > b$ then $\Delta'(x,a,b) < 0$. The second consequence is for quadratic potentials, if A is quadratic then $A'' = cte$ and this implies that $\Delta''(x,a,b) = 0$, thus $\Delta(x,a,b) = \Delta(0,a,b) + x\Delta'(0,a,b)$, for $x \in S^1$. The twist property give us a certain geometric structure on the family $S(x,a)$. If we assume a is optimal for $x = 0$ and define $a^- = \{w \in \Sigma,\ w < a\}$ and $a^+ = \{w \in \Sigma,\ w > a\}$ we get the picture described by Fig. 6.

Indeed, $\Delta(0,a,b) > 0$ and $\Delta'(x,a,b) > 0$ because $b > a$, thus $\Delta(x,a,b)$ is increasing what means that $S(x,a)$ and $S(x,b)$ has no intersection. On the other hand $c \in a^-$ which means that $\Delta'(x,a,b) < 0$ thus $S(x,c)$ can intersect $S(x,a)$ in just one point. Reciprocally, the twist property allow us to determinate the exact order of every three members a, b and c from the geometrical position of $S(x,a)$, $S(x,b)$

Fig. 6 $b \in a^+$ and $c \in a^-$

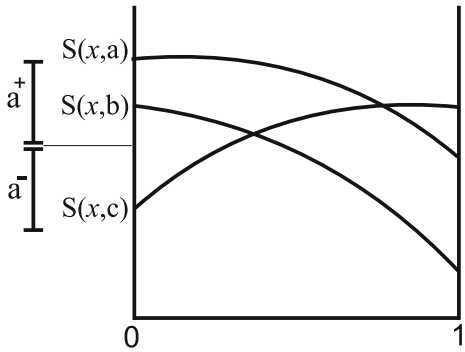
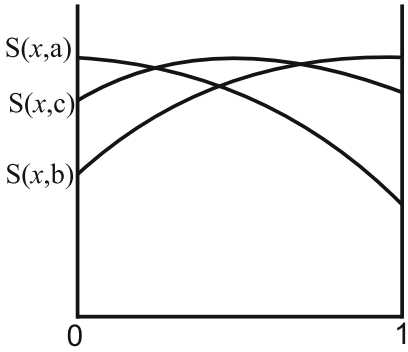


Fig. 7 Triangle property



and $S(x, c)$. We call this the triangle property; this means that if the corresponding positions are as in the Fig. 7, then, we get that $a > c > b$.

Proposition 2 *Suppose S satisfies the twist condition for some fixed a and b, the positions of the graphs of $S(., a)$ and $S(., b)$ are described by Fig. 5. We assume x_0 is such that $0 = \Delta(x_0, a, b) = S(x_0, a) - S(x_0, b)$, then $\frac{\partial S}{\partial x}(x_0, a) < \frac{\partial S}{\partial x}(x_0, b)$.*

Proof The proof follows from the fact that $\frac{\partial S}{\partial x}(x_0, a) - \frac{\partial S}{\partial x}(x_0, b) = \Delta'(x_0, a, b) < 0$.

The twist property assures a transversality condition on the intersections of the leaves described by the different graphs of $S(., a)$ (see beginning of Sect. 4 in [37]). We point out that the twist condition was not explicitly considered in [37].

4.1 Invariance Properties of the Envelope

We already know that $b(x)$ given by $b_{\lambda,A}(x) = \sup_{a \in \Sigma} S(x, a) = S(x, a(x))$, is the upper envelope of the family $S(x, a)$. We remind the reader that the map $\mathbb{T}^{-1}$ is defined by $\mathbb{T}^{-1}(x, a) = (\tau_{a_0}(x), \sigma(a))$. It is also well defined

$$S(x,a) = A(\tau_{a_0}(x)) + \lambda S(\tau_{a_0}(x), \sigma(a)).$$

We will prove that the upper envelope of the family $S(x,a)$ is invariant by $\mathbb{T}^{-1}$. Abusing of the notation we set $a(x)$ as ***the set of such solutions for a given fixed*** x; $a(x)$ is indeed a multi function, that is, $b(x) = S(x,a),\ \forall a \in a(x)$. We are going to prove that the first symbol in $a(x)$ uniquely determined by the symbol i_0 that turns out to be the maximum $b(x) = \max_i\{\lambda\, b(\tau_i x) + A(\tau_i x)\}$, more precisely, $b(x) = \lambda\, b(\tau_{i_0} x) + A(\tau_{i_0} x)$.

We begin with a technical and crucial lemma.

Lemma 1 *If A is twist and $a > b$, with $d(a,b) = \frac{1}{2^N}$ then the angle α between two intersecting $S(x,a) = S(x,b)$ satisfy*

$$\tan(\alpha) = \Delta'(x,a,b) \le \|A'\|_\infty \left(\frac{\lambda}{2}\right)^N \frac{2}{2-\lambda}.$$

Proof As A is twist $S(x,a)$ and $S(x,b)$ are transversal and the positive angle is given by $\tan(\alpha) = \Delta'(x,a,b)$.

$$\Delta'(x,a,b) = \frac{\partial S}{\partial x}(x,a) - \frac{\partial S}{\partial x}(x,b) = \sum_{k=0}^{\infty} \lambda^k (A'(\tau_{k,a}x) - A'(\tau_{k,b}x)) \frac{1}{2^{k+1}}$$

But

$$\frac{1}{2}\sum_{k=n}^{\infty}\left(\frac{\lambda}{2}\right)^k |A'(\tau_{k,a}x) - A'(\tau_{k,b}x)| \le \|A'\|_\infty \sum_{k=n}^{\infty}\left(\frac{\lambda}{2}\right)^k =$$

$$= \|A'\|_\infty \left(\frac{\lambda}{2}\right)^N \frac{1}{1-\frac{\lambda}{2}} = \|A'\|_\infty \left(\frac{\lambda}{2}\right)^N \frac{2}{2-\lambda}.$$

We point out that we do not need to take λ close to 1 for the above result. Now, we want to show that for any fixed λ, under the twist condition plus another technical condition, there exist a finite number of points c_j, $j = 1, 2, \ldots, k$, such that

$$b(x) = \sup_{c\in\{1,2,\ldots,d\}^{\mathbb{N}}} S(x,c) = \sup_{j=1,2,\ldots,k} S(x,c_j).$$

A natural question is to ask about the nature of these points c_j, $j = 1, 2, \ldots, k$. In the case the λ-maximizing probability is a unique periodic orbit and A is twist we will be able to describe some properties (see Sect. 6). Some properties depend of the combinatorics of the position of the orbits (see Theorem 4). It can happen (see example below) that the λ-maximizing probability is a periodic orbit of period 2 and we need to use 3 points c_1, c_2, c_3 in the above equation (see Fig. 11).

Lemma 2 *If $b(x) = \lambda\, b(\tau_{i_0} x) + A(\tau_{i_0} x)$ then $i_0 * a(\tau_{i_0} x) \in a(x)$, where $*$ means the concatenation. Reciprocally, if $b(x) = S(x,c)$, then, $b(\tau_{c_0} x) = S(\tau_{c_0} x, \sigma c)$, and*

$b(x) = \lambda\, b(\tau_{c_0} x) + A(\tau_{c_0} x)$. *In other words, if* $b(x) = S(x, c)$ *then the first symbol* $i = c_0$ *of* c *attains the supremum* $b(x) = \max_i \{\lambda\, b(\tau_i x) + A(\tau_i x)\}$.

Proof Suppose that $b(x) = \lambda\, b(\tau_{i_0} x) + A(\tau_{i_0} x)$ and $c \in a(\tau_{i_0} x)$, then $b(\tau_{i_0} x) = S(\tau_{i_0} x, c)$. An easy computation shows that $\lambda\, b(\tau_{i_0} x) + A(\tau_{i_0} x) = A(\tau_{i_0} x) + \lambda\, S(\tau_{i_0} x, c) = S(x, i_0 * c)$, so $b(x) = S(x, i_0 * c)$ which means that $i_0 * c \in a(x)$.

For the reciprocal suppose $b(x) = S(x, c) = A(\tau_{c_0} x) + \lambda S(\tau_{c_0} x, \sigma c)$. Since $b(x) = \max_i \{\lambda\, b(\tau_i x) + A(\tau_i x)\} \geq \lambda\, b(\tau_{c_0} x) + A(\tau_{c_0} x)$, we get,

$$A(\tau_{c_0} x) + \lambda S(\tau_{c_0} x, \sigma c) \geq \lambda\, b(\tau_{c_0} x) + A(\tau_{c_0} x),$$

which is equivalent to $b(\tau_{c_0} x) \leq S(\tau_{c_0} x, \sigma c)$, thus $b(\tau_{c_0} x) = S(\tau_{c_0} x, \sigma c)$. Substituting this in the previous equation we have that $b(x) = S(x, c) = A(\tau_{c_0} x) + \lambda b(\tau_{c_0} x)$ (Fig. 8).

If we suppose additionally that W satisfies the twist condition then, if $a(\tau_{i_0} x)$ is not a single point, then by Proposition 2, the function b is not differentiable at x because $b(x) = S(x, i_0 * c)$ and $b(x) = S(x, i_0 * d)$. However, $S'(x, i_0 * c) \neq S'(x, i_0 d)$, if $c \neq d$, where $c, d \in a(\tau_{i_0} x)$.

Corollary 1 *The set* $\Omega = \{(x, a) \in [0, 1] \times \Sigma \mid b(x) = S(x, a)\}$ *is* $\mathbb{T}^{-1}$*-invariant.*

Proof Indeed, if $(x, a) \in \Omega$ then $b(x) = S(x, a)$ and by Lemma 2 we have $b(\tau_{c_0} x) = S(\tau_{c_0} x, \sigma c)$. Thus $\mathbb{T}^{-1}(x, a) \in \Omega$.

Definition 8 A crossing point $x = x_{ab}$ is the single point x satisfying $S(x, a) = S(x, b)$ with $a > b$.

When A is twist, the crossing points are ordered according to the order of a, b and c as in the above figure.

Definition 9 A turning point is a point x such that $b(x) = \lambda\, b(\tau_{i_0} x) + A(\tau_{i_0} x)$ for more than one symbol i_0. The concept of turning point was introduced in [13, 30]. A turning point is **simple** if its forward orbit is finite.

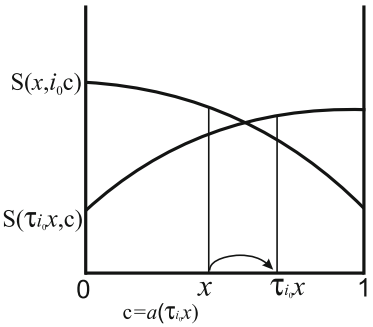


Fig. 8 Invariance of the boundary

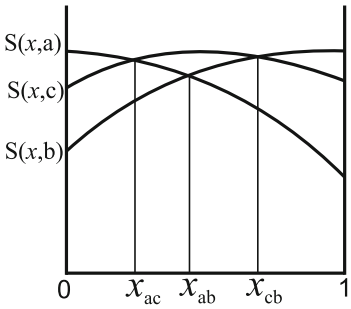


Fig. 9 Ordered intersections between different $S(x, c)$

Corollary 2 *Assume A satisfies the twist condition and moreover that there exists a finite number of turning points and that each one is simple, then the boundary of the attractor is given by a finite number of C^2 pieces (of unstable manifolds).*

Proof Let x be a point on the boundary of the attractor where the optimal symbolic changes, see Fig. 11. If $S(x, a) = b(x) = S(x, c)$ and $a_i = c_i$ for $i = 0, \ldots, N-1$ we get from Lemma 2 that $b(x) = S(x, c)$, then, $b(\tau_{c_0} x) = S(\tau_{c_0} x, \sigma c)$ and $b(x) = S(x, a)$, then, $b(\tau_{a_0} x) = S(\tau_{a_0} x, \sigma a)$. Choosing $x_1 = \tau_{a_0} x$ we get $S(x_1, \sigma a) = S(x_1, \sigma c)$. Proceeding in this way we obtain $S(x_{N-1}, \sigma^{N-1} a) = S(x_{N-1}, \sigma^{N-1} c)$. What means that $z = x_{N-1}$ is a turning point and $T^N(z) = x$. In this way, we conclude that any point x such that $S(x, a) = b(x) = S(x, c)$ lies in the orbit of a turning point. Since the number of turning points is finite and its orbits are finite, because they are simple, we obtain that there is just a finite number of this points. Finally, by the twist property we guarantee that $\sharp\{x \mid S(x, a) = b(x) = S(x, c)\} = 1$ that is, the number of pieces in the boundary is finite (Fig. 9).

We will present later examples where b is explicit and have a finite number o realizers. Therefore the boundary of the attractor is given by a finite number of C^2 pieces (see Sect. 7.4).

In general, explicit computations are very difficult to find, but we will present some computational evidence to illustrate the conclusion of Corollary 2.

Example 1 Take $\varepsilon = 0.005$, $\lambda = 0.51$, $drift = 0.05$ and $gap = 0.001$, we use a truncated version $S(x, a) = \sum_{k=0}^{7} \lambda^k A(\tau_{k,a} x)$ where $A(x) := A_\varepsilon(x) = -(1.010\ x - 0.455)^2$ is a perturbation of $-(x - 0.5)^2$ by $A_\varepsilon(x) = -(x - 0.5 + \varepsilon\phi(x) + drift)^2$ and $\phi(x) = 2x - 1$. The figure below shows the maximum $b(x) = \max_a S(x, a)$ in a grid of 25 divisions of [0, 1], and suggest the form of the graph of b in the Fig. 10. This figure suggest that there is 3 pieces $S(x, 10101....)$, $S(x, 01010....)$ and a unknown $S(x, c_0 c_1 c_2 ...)$. Since the perturbed potential still having the twist property we get $c_0 = 0$. Taking x in the right side of de second crossing point v we get $b(x) = S(x, c_0 * \sigma c)$ and from Corollary 1 we get $b(\tau_{c_0} x) = S(\tau_{c_0} x, \sigma c)$, in particular $b(\tau_{c_0} x) = b(\tau_0 x)$ lies in the right side of the first crossing point u because the first symbol of the optimal sequences for points before u is 1. Therefore, σc should be (0101...). From this hypothetic deductions we can suppose that, if u and v are the crossing points, then the formula for the superior envelop $b(x)$ should be

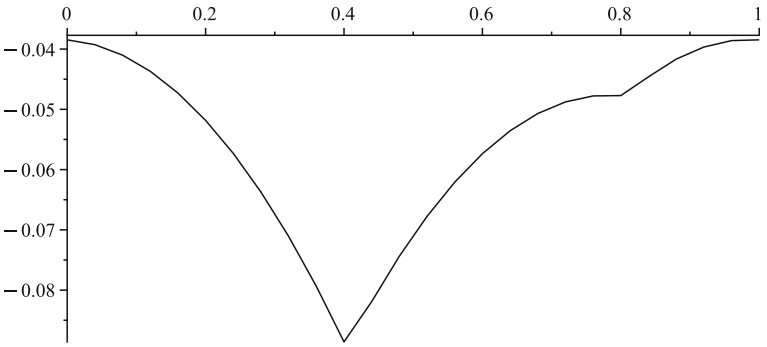


Fig. 10 The graph of $b(x)$

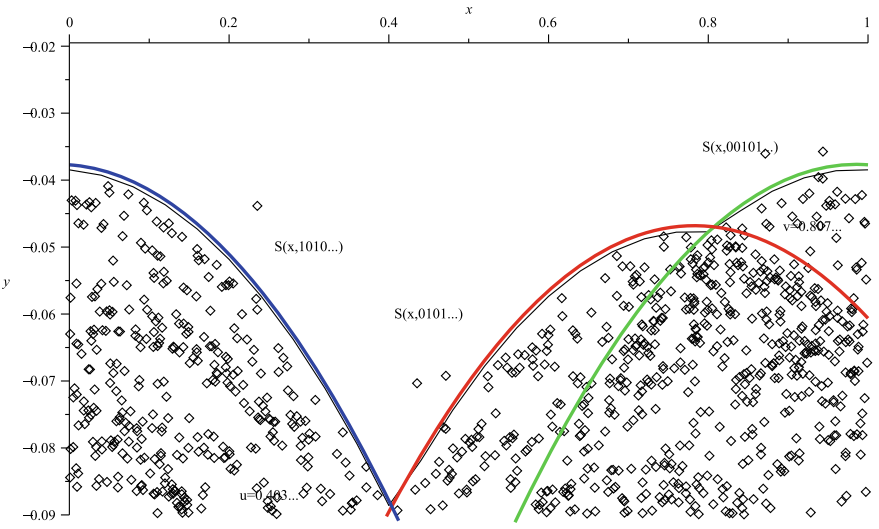


Fig. 11 Iteration by 4000 times of F

$$b(x) := \begin{cases} S(x, 101010....) \ , \ 0 < x \leq u \\ S(x, 010101....) \ , \ u < x \leq v \\ S(x, 001010....) \ , \ v < x \leq 1 \end{cases}$$

This is exactly what the Fig. 11 shows, that is, we already know from Corollary 2 that there is only a finite number of pieces, we just deduce now what are the geometric positions of these pieces. In the graph we plot $b(x) + gap = b(x) + 0.001$ in order to distinguish the difference between this and the picture computationally obtained. If we iterate some orbits close to the attractor by the transformation $F(x, s) = (T(x), \lambda s + A(x))$, we can see that there is numerical evidence that our claim is true in this particular case.

Denote by $\mathscr{T}$ the set of turning points and $\Lambda = \cup_{n\geq 0} T^n(\mathscr{T})$. In order to characterize the turning points we follow a discounted version of the notation introduced by [9] for Sturmian measures when the symbols are $\{0, 1\}$.

Definition 10 If $b(x) = \max_i\{\lambda\, b(\tau_i x) + A(\tau_i x)\}$, we define the remainders associated with the b and A as $r(x, a) = b(x) - \lambda\, b(\tau_{a_0} x) - A(\tau_{a_0} x)$,

$$R(x) = r(x, 0\cdots) - r(x, 1\cdots) = (\lambda\, b(\tau_1 x) + A(\tau_1 x)) - (\lambda\, b(\tau_0 x) + A(\tau_0 x)).$$

So, $r(x, a) \geq 0$ and attains zero with the right symbol a_0. Also, $R(x) = 0$, if and only if, x is a turning point that is, $\mathscr{T} = R^{-1}(0)$.

Definition 11 A continuous potential A satisfy the k-Sturmian condition if $\sharp R^{-1}(0) = k$. In particular, there is just k turning points.

In [9] Sturmian measures are that ones where $k = 1$. In a slightly different setting the author shows that $A(x) = cos(2\pi(x - \omega))$ satisfy the 1-Sturmian condition for any $\omega \in \mathbb{R}/\mathbb{Z}$.

Lemma 3 *Λ contains the set of points x where $a(x)$ is not a single point.*

Proof Suppose that there is two different elements $c, d \in a(x)$, where c and d are of the form $c = (i_0, i_1, \ldots, i_{n-1}, 1, c_{n+1}, ...)$, $d = (i_0, i_1, \ldots, i_{n-1}, 0, d_{n+1}, \ldots)$. Take $z = \tau_{i_{n-1}} \cdots \tau_{i_0} x$. Applying Corollary 1 we get:

$$c \in a(x) \Rightarrow b(x) = S(x, c) \Rightarrow b(\tau_{i_0} x) = S(\tau_{i_0} x, \sigma c) \cdots$$

$$\Rightarrow b(\tau_{i_1}\tau_{i_0} x) = S(\tau_{i_1}\tau_{i_0} x, \sigma^2 c) \cdots \Rightarrow b(z) = S(z, (1, c_{n+1}, \ldots)).$$

$$d \in a(x) \Rightarrow b(x) = S(x, d) \Rightarrow b(\tau_{i_0} x) = S(\tau_{i_0} x, \sigma d) \cdots \Rightarrow$$

$$\Rightarrow b(\tau_{i_1}\tau_{i_0} x) = S(\tau_{i_1}\tau_{i_0} x, \sigma^2 d) \cdots \Rightarrow b(z) = S(z, (0, d_{n+1}, \ldots)).$$

Thus, $b(z) = \lambda\, b(\tau_0 z) + A(\tau_0 z)$ and $b(z) = \lambda\, b(\tau_1 z) + A(\tau_1 z)$, by Lemma 2, that is, $z \in \mathscr{T}$. Since $T^n(z) = x$ we get $x \in \Lambda$.

If the turning points are finite (A satisfying k-Sturmian condition) and pre-periodic points for T, then there exists finitely many points where the optimal symbolic changes because $\sharp\Lambda < \infty$.

Corollary 3 *If $\sharp\Lambda$ is finite then the graph of b is a union of a finite number of $S(x, a)$.*

Proof The claim follows from Lemma 3.

5 Symmetric Twist Potentials

In this section we exhibit some explicit examples.

Theorem 3 *Let A be a symmetric potential, that is, $A(1-x) = A(x)$ for any $x \in [0,1]$. In addition we assume that A is twist. Denote by $b : [0,1] \to \mathbb{R}$ the function such that*

$$b(x) = \begin{cases} S(x, (10)^\infty), \ 0 \leq x \leq 1/2; \\ S(x, (01)^\infty), \ 1/2 < x \leq 1. \end{cases}$$

Then, b is a λ-calibrated subaction for $A(x)$, that is, for any $x \in [0,1]$

$$b(x) = \max_{i=0,1} \{\lambda \, b(\tau_i x) + A(\tau_i x)\}.$$

Proof As A is symmetric $A(1/2 - t) = A(1/2 + t)$ for $t \in [0, 1/2]$. We claim that $S(x, (10)^\infty) = S((1-x), (01)^\infty)$. Indeed

$$\tau_0(1-x) = \frac{1-x}{2} = 1/2 - x/2 \text{ and } \tau_1(x) = \frac{1+x}{2} = 1/2 + x/2,$$

then, $A(\tau_0(1-x)) = A(\tau_1(x))$. Analogously, $\tau_1\tau_0(1-x) = 1/4 - x/4 + 1/2$ and $\tau_0\tau_1(x) = \frac{1+x}{2} = 1/4 + x/4 = 1/2 - (1/4 - x/4)$, then $A(\tau_1\tau_0(1-x)) = A(\tau_0\tau_1(x))$, and so on. Thus

$$S(x, (10)^\infty) = A(\tau_1(x)) + \lambda \, A(\tau_0 \circ \tau_1(x)) + \lambda^2 \, A(\tau_1 \circ \tau_0 \circ \tau_1(x)) + \cdots =$$

$$A(\tau_0(1-x)) + \lambda \, A(\tau_1\tau_0(1-x)) + \lambda^2 \, A(\tau_0 \circ \tau_1 \circ \tau_0(1-x)) + \cdots = S((1-x), (01)^\infty).$$

In particular, $S(0, (10)^\infty) = S(1, (01)^\infty)$ and $S(1/2, (10)^\infty) = S(1/2, (01)^\infty)$, that is $b(x)$ is continuous. By the twist property $S(x, (10)^\infty)$ and $S(x, (01)^\infty)$ are transversal in $x = 1/2$, then, as can not exist two points of intersection we get

$$S(x, (10)^\infty) > S(x, (01)^\infty) \text{ if } x < 1/2;$$
$$S(x, (10)^\infty) < S(x, (01)^\infty) \text{ if } x > 1/2.$$

Now we will prove that the above b is a λ-calibrated subaction for $A(x)$.

We divide the argument in two cases:

Case 1- $x < 1/2$

If $i = 0$ then

$$\begin{aligned} \lambda b(\tau_i x) + A(\tau_i x) &= A(\tau_0 x) + \lambda b(\tau_0 x) \\ &= A(\tau_0 x) + \lambda S(\tau_0 x, (10)^\infty) \\ &= S(x, (01)^\infty) < S(x, (10)^\infty) \end{aligned}$$

because $x < 1/2$.

If $i = 1$ then

$$\begin{aligned}\lambda b(\tau_i x) + A(\tau_i x) &= A(\tau_1 x) + \lambda b(\tau_1 x)\\ &= A(\tau_1 x) + \lambda S(\tau_1 x, (01)^\infty)\\ &= S(x, (10)^\infty)\end{aligned}$$

because $\tau_1 x > 1/2$.
Thus, $\max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\} = S(x, (10)^\infty) = b(x)$ if $x < 1/2$.
Case 2- $x > 1/2$
If $i = 0$ then

$$\begin{aligned}\lambda b(\tau_i x) + A(\tau_i x) &= A(\tau_0 x) + \lambda b(\tau_0 x)\\ &= A(\tau_0 x) + \lambda S(\tau_0 x, (10)^\infty)\\ &= S(x, (01)^\infty)\end{aligned}$$

because $\tau_0 x < 1/2$. If $i = 1$ then

$$\begin{aligned}\lambda b(\tau_i x) + A(\tau_i x) &= A(\tau_1 x) + \lambda b(\tau_1 x)\\ &= A(\tau_1 x) + \lambda S(\tau_1 x, (01)^\infty)\\ &= S(x, (10)^\infty) < S(x, (01)^\infty)\end{aligned}$$

because $x > 1/2$. Thus, $\max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\} = S(x, (01)^\infty) = b(x)$ if $x > 1/2$.

It follows from above that in this case the λ-calibrated subaction is piecewise differentiable if A is differentiable. It is piecewise analytic (two domains of analyticity) if A is analytic.

6 A characterization of When the Boundary is Piecewise Smooth in the Case of Period 2 and 3

We will present a characterization of when the boundary is piecewise smooth in the case the λ-maximizing probability has period 2 and 3 (see Theorem 4). As we know that a λ-calibrated subaction for $A(x)$, that is, for any $x \in [0, 1]$ $b(x) = \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}$, is unique and $\bar{b}(x) = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(x, a)$, is also a solution of this equation, we get that the superior envelope of $\{S(x, a) \mid a \in \{0, 1\}^{\mathbb{N}}\}$ is piecewise regular as much as A.

Lemma 4 *If* $0 < x < y < 1$ *and* $S(x, a') = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(x, a)$ *and* $S(y, a'') = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(y, a)$ *then there is* $x < z < y$ *such that* $S(z, a') = S(z, a'')$. *In particular, from the twist property,* $a' > a''$ *and*

$$\begin{aligned}&S(x, a') > S(x, a'') \text{ if } x < z;\\ &S(x, a') < S(x, a'') \text{ if } x > z.\end{aligned}$$

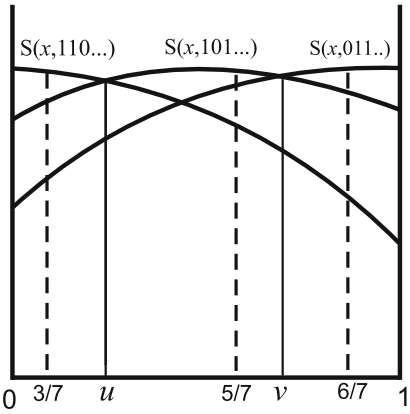


Fig. 12 Ordering the intersections according to a periodic orbit

We assume here that $T(x)$ is the transformation $2\,x \pmod 1$. Now we going to consider the case where the maximizing measure is supported in a periodic orbit of period 3. We know that if the minimum period is 3 there is just two possible periodic sequences: $(100100100...) = (100)^\infty$ and $(110110110...) = (110)^\infty$. We choose the case $(110110110...) = (110)^\infty$ with the correspondent periodic point $x_0 = 3/7 < T^2(x_0) = 5/7 < T(x_0) = 6/7$ (Fig. 12).

Theorem 4 *Let A be a twist potential such that*
$S(x_0, (110)^\infty) = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(x_0, a)$, $S(T(x_0), (011)^\infty) = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(T(x_0), a)$,
$S(T^2(x_0), (101)^\infty) = \sup_{a\in\{0,1\}^{\mathbb{N}}} S(T^2(x_0), a)$, *where* $T^3(x_0) = x_0$. *Let* $u \in [x_0, T^2(x_0)]$ *and* $v \in [T^2(x_0), T(x_0)]$ *given by Lemma 4, that is,* $S(u, (110)^\infty) = S(u, (101)^\infty)$ *and* $S(v, (101)^\infty) = S(v, (011)^\infty)$. *Denote by* $b : [0, 1] \to \mathbb{R}$ *the function such that*

$$b(x) = \begin{cases} S(x, (110)^\infty),\ 0 \le x \le u; \\ S(x, (101)^\infty),\ u \le x \le v; \\ S(x, (011)^\infty),\ v \le x \le 1. \end{cases}$$

Then, b is a λ-calibrated subaction for $A(x)$, that is, for any $x \in [0, 1]$ $b(x) = \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}$, if and only if, $\tau_1[0, u] \subseteq [u, v]$, $\tau_1[u, v] \subseteq [v, 1]$ and $\tau_0[v, 1] \subseteq [0, u]$.

Proof We must to divide in several cases.
Case 1: Consider $0 \le x \le u$.
As $\tau_0(x) = 1/2x < u$ thus

$$\lambda b(\tau_0 x) + A(\tau_0 x) = \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) =$$

$$= \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) = S(x, 0(110)^\infty) =$$

$$= S(x, (011)^\infty) < S(x, (110)^\infty) = b(x)$$

As $\tau_1(x) = 1/2x + 1/2 > u$ we have two possibilities
(a) If $\tau_1(x) \in (u, v]$ then

$$\lambda b(\tau_1 x) + A(\tau_1 x) = \lambda S(\tau_1 x, (101)^\infty) + A(\tau_1 x) =$$

$$= \lambda S(\tau_1 x, (101)^\infty) + A(\tau_1 x) = S(x, 1(101)^\infty) =$$

$$= S(x, (110)^\infty) = b(x)$$

(b) If $\tau_1(x) \in [v, 1]$ then

$$\lambda b(\tau_1 x) + A(\tau_1 x) = \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) =$$

$$= \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) = S(x, 1(011)^\infty) =$$

$$= S(x, (101)^\infty) < S(x, (110)^\infty) = b(x).$$

Thus

$$b(x) \begin{cases} = \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ if } \tau_1[0, u] \subseteq [u, v] \\ < \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ otherwise.} \end{cases}$$

for $0 \leq x \leq u$.

Case 2: Consider $u \leq x \leq v$.

As $\tau_0(x) = 1/2x < v$ we have two possibilities
(a) If $\tau_0(x) \in [0, u]$ then

$$\lambda b(\tau_0 x) + A(\tau_0 x) = \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) =$$

$$= \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) = S(x, 0(110)^\infty) =$$

$$= S(x, (011)^\infty) < S(x, (101)^\infty) = b(x)$$

(b) If $\tau_0(x) \in [u, v]$ then

$$\lambda b(\tau_0 x) + A(\tau_0 x) = \lambda S(\tau_0 x, (101)^\infty) + A(\tau_0 x) =$$

$$= \lambda S(\tau_0 x, (101)^\infty) + A(\tau_0 x) = S(x, 0(101)^\infty) =$$

$$= S(x, 0(101)^\infty) < S(x, (101)^\infty) = b(x),$$

because $S(x, 0(101)^\infty) > S(x, (101)^\infty)$ contradicts the twist condition, as one can see from the Fig. 13. As $\tau_1(x) \in [5/7, 6/7]$ we have two possibilities

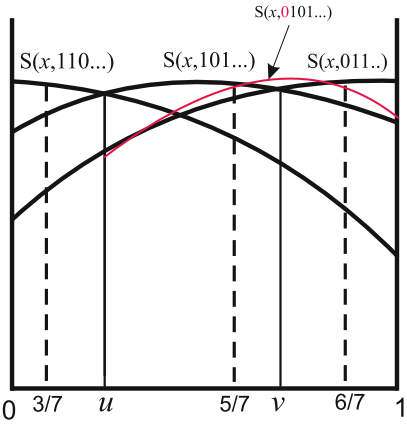


Fig. 13 Avoiding the wrong symbolic $S(x, 0101...)$

(a) If $\tau_1(x) \in [u, v]$ then

$$\lambda b(\tau_1 x) + A(\tau_1 x) = \lambda S(\tau_1 x, (101)^\infty) + A(\tau_1 x) =$$

$$= \lambda S(\tau_1 x, (101)^\infty) + A(\tau_1 x) = S(x, 1(101)^\infty) =$$

$$= S(x, (110)^\infty) < S(x, (101)^\infty) = b(x).$$

(b) If $\tau_1(x) \in [v, 1]$ then

$$\lambda b(\tau_1 x) + A(\tau_1 x) = \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) =$$

$$= \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) = S(x, 1(011)^\infty) =$$

$$= S(x, (101)^\infty) = b(x).$$

Thus

$$b(x) \begin{cases} = \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ if } \tau_1[u, v] \subseteq [v, 1] \\ < \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ otherwise.} \end{cases}$$

for $u \leq x \leq v$.

Case 3: Consider $v \leq x \leq 1$.
As $\tau_0(x) = 1/2x < 1/2$ we have two possibilities
(a) If $\tau_0(x) \in [0, u]$ then

$$\lambda b(\tau_0 x) + A(\tau_0 x) = \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) =$$

$$= \lambda S(\tau_0 x, (110)^\infty) + A(\tau_0 x) = S(x, 0(110)^\infty) =$$

$$= S(x, (011)^\infty) = b(x).$$

(b) If $\tau_0(x) \in [u, v]$ then

$$\lambda b(\tau_0 x) + A(\tau_0 x) = \lambda S(\tau_0 x, (101)^\infty) + A(\tau_0 x) =$$

$$= \lambda S(\tau_0 x, (101)^\infty) + A(\tau_0 x) = S(x, 0(101)^\infty) =$$

$$= S(x, 0(101)^\infty) < S(x, (011)^\infty) = b(x),$$

because $S(x, 0(101)^\infty) > S(x, (011)^\infty)$ contradicts the twist (as one can see from Fig. 13).

As $\tau_1(x) \in [v, 1]$ we have

$$\lambda b(\tau_1 x) + A(\tau_1 x) = \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) =$$

$$= \lambda S(\tau_1 x, (011)^\infty) + A(\tau_1 x) = S(x, 1(011)^\infty) =$$

$$= S(x, (101)^\infty) < S(x, (011)^\infty) = b(x).$$

Thus

$$b(x) \begin{cases} = \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ if } \tau_0[v, 1] \subseteq [0, u] \\ < \max_{i=0,1}\{\lambda\, b(\tau_i x) + A(\tau_i x)\}, \text{ otherwise.} \end{cases}$$

for $v \leq x \leq 1$.

The characterization in the case the maximizing measure has support in an orbit of period n is similar. One needs to know the combinatorics of the position of the different points of the orbit and then proceed in an analogous way as in the case of period 3. We left this to the reader.

7 Twist Properties in the Case $T(x) = 2x \pmod 1$

Let us fix $a = (a_0, a_1, ...) \in \{0, 1\}^{\mathbb{N}}$. If A is differentiable we can differentiate S with respect to x

$$\frac{\partial S}{\partial x}(x, a) = \sum_{k=0}^{\infty} \lambda^k A'(\tau_{k,a} x) \frac{\partial}{\partial x} \tau_{k,a} x.$$

We observe that $\tau_{k,a} x$ has an explicit expression: $\tau_{k,a} x = \frac{1}{2^{k+1}} x + \psi_k(a)$, where

$$\psi_k(a) = \frac{a_0}{2^{k+1}} + \frac{a_1}{2^k} + \cdots + \frac{a_k}{2},$$

satisfy the recurrence relation $2\psi_{k+1}(a) = \psi_k(a) + a_{k+1}$. Thus

$$\frac{\partial S}{\partial x}(x,a) = \sum_{k=0}^{\infty} \lambda^k A'(\tau_{k,a}x)\frac{1}{2^{k+1}} = \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k A'(\tau_{k,a}x).$$

Analogously,

$$\frac{\partial^2 S}{\partial x^2}(x,a) = \sum_{k=0}^{\infty} \lambda^k A''(\tau_{k,a}x)\frac{1}{2^{k+1}}^2 = \frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{\lambda}{4}\right)^k A''(\tau_{k,a}x),$$

in particular, if $A'' < 0$ then $\frac{\partial^2 S}{\partial x^2}(x,a) < 0, \forall a \in \Sigma$. Even if A is not C^2 we have the concavity of S from A:

Lemma 5 *Let A be a C^0 potential in S^1.*
If A is concave (strictly) then $S(x,a)$ is concave (strictly), $\forall a \in \Sigma$.

Proof Fixed $a \in \Sigma$ consider $x < y$ and $t \in [0,1]$ then

$$S((1-t)x + ty, a) = \sum_{k=0}^{\infty} \lambda^k A(\tau_{k,a}[(1-t)x + ty]).$$

Since $\tau_{k,a}(1-t)x + ty = \frac{1}{2^{k+1}}[(1-t)x + ty] + \psi_k(a) = (1-t)\tau_{k,a}x + t\tau_{k,a}y$, we get

$$S((1-t)x + ty, a) = \sum_{k=0}^{\infty} \lambda^k A((1-t)\tau_{k,a}x + t\tau_{k,a}y) \geq$$

$$\geq \sum_{k=0}^{\infty} \lambda^k [(1-t)A(\tau_{k,a}x) + tA(\tau_{k,a}y)] = (1-t)S(x,a) + tS(y,a).$$

7.1 Formal Computations

First we prove two technical lemmas about recursive sums.

Lemma 6 *Let $\psi_k(a)$ be the function defined above, then* $\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k \psi_k(a) = \frac{2}{4-\lambda} Z(a)$, *where* $Z(a) = \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k$.

Proof Consider $H = \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k \psi_k(a)$ then

$$
\begin{array}{ll}
2H & = 2\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k \psi_k(a) \qquad\qquad = 2\psi_0(a) + \sum_{k=1}^{\infty}\left(\frac{\lambda}{2}\right)^k 2\psi_k(a) \\
& = a_0 + \sum_{k=1}^{\infty}\left(\frac{\lambda}{2}\right)^k [\psi_{k-1}(a) + a_k] \quad = a_0 + \sum_{k=1}^{\infty}\left(\frac{\lambda}{2}\right)^k \psi_{k-1}(a) + \sum_{k=1}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k \\
& = \frac{\lambda}{2}\sum_{k=1}^{\infty}\left(\frac{\lambda}{2}\right)^{k-1} \psi_{k-1}(a) + \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k \quad = \frac{\lambda}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k \psi_k(a) + \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k \\
& = \frac{\lambda}{2}H + Z(a).
\end{array}
$$

Thus, $H = \frac{2}{4-\lambda} Z(a)$, where $Z(a) = \displaystyle\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k$ is the expansion in the bases $\frac{2}{\lambda}$ of the number $Z(a)$.

Lemma 7 *If $\lambda < 1$ the function $Z : \Sigma \to [0, 1]$ given by $Z(a) = \displaystyle\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k a_k$ is strictly increasing with respect to the lexicographical order. In particular, if $b > a$ then*

$$Z(b) - Z(a) \geq \left(\frac{\lambda}{2}\right)^n \left(\frac{1-\lambda}{1-\frac{\lambda}{2}}\right),$$

where n is the first digit where a is different from b.

Proof Take $a = (i_0, \ldots, i_{n-1}, 0, a_{n+1}, i_{n+2}, ...) < b = (i_0, \ldots, i_{n-1}, 1, b_{n+1}, b_{n+2}, ...)$, then,

$$Z(b) - Z(a) = \left(\frac{\lambda}{2}\right)^n (1-0) + \sum_{k=n+1}^{\infty}\left(\frac{\lambda}{2}\right)^k (b_k - a_k) \geq$$

$$\geq \left(\frac{\lambda}{2}\right)^n - \sum_{k=n+1}^{\infty}\left(\frac{\lambda}{2}\right)^k = \left(\frac{\lambda}{2}\right)^n - \frac{(\frac{\lambda}{2})^{n+1}}{1-\frac{\lambda}{2}} =$$

$$\left(\frac{\lambda}{2}\right)^n \left(1 - \frac{\frac{\lambda}{2}}{1-\frac{\lambda}{2}}\right) = \left(\frac{\lambda}{2}\right)^n \left(\frac{1-\lambda}{1-\frac{\lambda}{2}}\right) > 0$$

We are going now to compute $\frac{\partial S}{\partial x}(x, a)$ for x^m for $m = 0, 1, 2$.[1]
(a) $A(x) = 1$
In that case, $S_1(x, a) = \sum_{k=0}^{\infty} \lambda^k 1 = \frac{1}{1-\lambda}$, so $\frac{\partial S}{\partial x}(x, a) = 0$.
(b) $A(x) = x$
In that case,
$S_x(x, a) = \sum_{k=0}^{\infty} \lambda^k \tau_{k,a} x$, so $\frac{\partial S}{\partial x}(x, a) = \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k = \frac{1}{2-\lambda}$.
(c) $A(x) = x^2$
In that case,
$S_{x^2}(x, a) = \sum_{k=0}^{\infty} \lambda^k (\tau_{k,a} x)^2$,

[1] This potentials are actually defined in $\mathbb{R}$ because they are not continuous functions on S^1, but some combination of $1, x, x^2, \ldots$ allow us to build an 1-periodic function.

so $\frac{\partial S}{\partial x}(x,a) = \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k 2(\tau_{k,a}x) = \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k(\tau_{k,a}x)$. Thus,

$$\begin{aligned}\frac{\partial S}{\partial x}(x,a) &= \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k\left[\frac{1}{2^{k+1}}x + \psi_k(a)\right]\\ &= \frac{x}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda}{4}\right)^k + \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k\psi_k(a)\\ &= \frac{2}{4-\lambda}x + \sum_{k=0}^{\infty}\left(\frac{\lambda}{2}\right)^k\psi_k(a)\end{aligned}$$

Applying Lemma 6 we have $\frac{\partial S_{x^2}}{\partial x}(x,a) = \frac{2}{4-\lambda}x + \frac{2}{4-\lambda}Z(a)$.

Theorem 5 *If $A(x) = c_0 + c_1x + c_2x^2$ is 1-periodic differentiable in $S^1 - \{0\}$ then*

$$\frac{\partial S}{\partial x}(x,a) = \left(\frac{c_1}{2-\lambda} + \frac{2c_2}{4-\lambda}x\right) + \frac{2c_2}{4-\lambda}Z(a).$$

Moreover, A is twist if and only if $c_2 < 0$.

Proof Using the notation $S_A(x,a) = \sum_{k=0}^{\infty}\lambda^k A(\tau_{k,a}x)$, one can easily show that S depends linearly of A. So if we have, $A(x) = c_0 + c_1x + c_2x^2$, then

$$S_A(x,a) = c_0 \cdot S_1(x,a) + c_1S_x(x,a) + c_2S_{x^2}(x,a).$$

We also can compute $\frac{\partial S}{\partial x}(x,a)$,

$$\frac{\partial S}{\partial x}(x,a) = c_0 \cdot 0 + c_1\frac{1}{2-\lambda} + c_2\left(\frac{2}{4-\lambda}x + \frac{2}{4-\lambda}Z(a)\right),$$

or

$$\frac{\partial S}{\partial x}(x,a) = \left(\frac{c_1}{2-\lambda} + \frac{2c_2}{4-\lambda}x\right) + \frac{2c_2}{4-\lambda}Z(a).$$

Moreover,

$$\frac{\partial S}{\partial x}(x,a) - \frac{\partial S}{\partial x}(x,a) = \frac{2c_2}{4-\lambda}(Z(a) - Z(b)).$$

Remember that, if $a > b$ then $Z(a) - Z(b) > 0$ (by Lemma 7). In this way, $\frac{\partial S}{\partial x}(x,a) - \frac{\partial S}{\partial x}(x,a) < 0$, if and only, if $c_2 < 0$.

Suppose that A is such that the λ-maximizing probability has period 2 and the subaction b_λ is the envelope of $S(x,(0,1,0,1,\ldots))$ and $S(x,(1,0,1,0\ldots))$. In this case, in order to get explicit examples of the associated b_λ, it is quite useful to have the explicit expression for the associated pre-orbits.

(a) The expression of the preimages using $(0,1,0,1\ldots)$. These are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{5}{8}$, $\frac{5}{16}$ .... We say that $\frac{1}{2}$ is the 0 level, $\frac{1}{4}$ is the 1 level , $\frac{5}{8}$ is the 2 level, and so on. One can show that, if m is even level, then $\frac{2^{m+2}-1}{3\,2^{m+1}}$. If m is in odd level the value it is the last one divided by 2.

(b) The expression of the preimages using $(1, 0, 1, 0, \ldots)$. These are $\frac{1}{2}$, $\frac{3}{4}$, $\frac{3}{8}$, $\frac{11}{16}$ .... We say $\frac{1}{2}$ is the 0 level, $\frac{3}{4}$ is the 1 level, $\frac{3}{8}$ is the 2 level, and so on. One can show that, if m is in odd level, then $\frac{2^{m+2}+1}{3\,2^{m+1}}$. If m is in even level the value it is the next one multiplied by 2.

7.2 A Special Quadratic Case

As an example let us consider $A(x) = -(x - 1/2)^2 = -1/4 + x - x^2$, then in this case $c_1 = 1$ and $c_2 = -1$

$$\frac{\partial S}{\partial x}(x, a) = \left(\frac{1}{2-\lambda} - \frac{2}{4-\lambda}x\right) - \frac{2}{4-\lambda}Z(a),$$

in particular

$$S(x, a) = S(0, a) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{2x}{4-\lambda}Z(a).$$

Thus,

$$\Delta(x, a, b) = S(x, a) - S(x, b) = \Delta(0, a, b) - \frac{2x}{4-\lambda}(Z(a) - Z(b))$$

$$\Delta'(x, a, b) = S'(x, a) - S'(x, b) = -\frac{2}{4-\lambda}(Z(a) - Z(b)).$$

This proves that $A(x) = -(x - 1/2)^2 = -1/4 + x - x^2$ is twist. Indeed, if $a > b$ then $Z(a) > Z(b)$ and so $\Delta'(x, a, b) < 0$. Note that when $\lambda \to 1$ the angles remain bounded away from zero.

7.3 Crossing Points for Quadratic Potentials

For the case $A(x) = -(x - 1/2)^2$ we can compute explicitly the crossing points $x_{ab} = x$ or equivalently $\Delta(x, a, b) = 0$, that is,

$$x_{ab} = \frac{4-\lambda}{2}\frac{\Delta(0, a, b)}{Z(a) - Z(b)}.$$

7.4 The Explicit λ-Calibrated Subaction for $A(x) = -(x - 1/2)^2$

Remember that $Z(a) = \sum_{k=0}^{\infty} \left(\frac{\lambda}{2}\right)^k a_k$. Note that $Z((01)^\infty) = 0 + \frac{\lambda}{2} + 0 + (\frac{\lambda}{2})^3 + 0 + \cdots = \frac{\lambda}{2}\frac{4}{4-\lambda^2}$. Moreover, $Z((10)^\infty) = 1 + 0 + (\frac{\lambda}{2})^2 + 0 + (\frac{\lambda}{2})^4 + 0 + \cdots = \frac{4}{4-\lambda^2}$. Note that

$$S(x, (01)^\infty) = S(0, (01)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{2x}{4-\lambda}Z((01)^\infty) =$$

$$S(0, (01)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{2x}{4-\lambda}\frac{\lambda}{2}\frac{4}{4-\lambda^2} =$$

$$S(0, (01)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{4\,\lambda\,x}{(4-\lambda)\,(4-\lambda^2)} =$$

$$S(0, (01)^\infty) + \frac{(8 - 2\,\lambda - \lambda^2)\,x}{(4-\lambda)\,(4-\lambda^2)} - \frac{1}{4-\lambda}x^2,$$

and

$$S(x, (10)^\infty) = S(0, (10)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{2x}{4-\lambda}Z((10)^\infty) =$$

$$S(0, (10)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{2x}{4-\lambda}\frac{4}{4-\lambda^2} =$$

$$S(0, (10)^\infty) + \left(\frac{1}{2-\lambda}x - \frac{1}{4-\lambda}x^2\right) - \frac{8\,x}{(4-\lambda)\,(4-\lambda^2)} =$$

$$S(0, (10)^\infty) + \frac{(2\,\lambda - \lambda^2)\,x}{(4-\lambda)\,(4-\lambda^2)} - \frac{1}{4-\lambda}x^2.$$

The value $S(0, (10)^\infty)$ will be explicitly obtained in the next proposition.

Observe that $S(1, (01)^\infty) = S(0, (01)^\infty) + \frac{(4-2\,\lambda)}{(4-\lambda)\,(4-\lambda^2)}$, and $S(1, (10)^\infty) = S(0, (10)^\infty) + \frac{(2\,\lambda-4)}{(4-\lambda)\,(4-\lambda^2)}$. By symmetry (see next proposition) we have that $S(1/2, (10)^\infty) = S(1/2, (01)^\infty)$, and therefore

$$S(1/2, (01)^\infty) = S(0, (01)^\infty) + \frac{6+\lambda}{4\,(4-\lambda)\,(2+\lambda)} =$$

$$S(1/2, (10)^\infty) = S(0, (10)^\infty) - \frac{2}{4\,(4-\lambda)\,(2+\lambda)}. \tag{8}$$

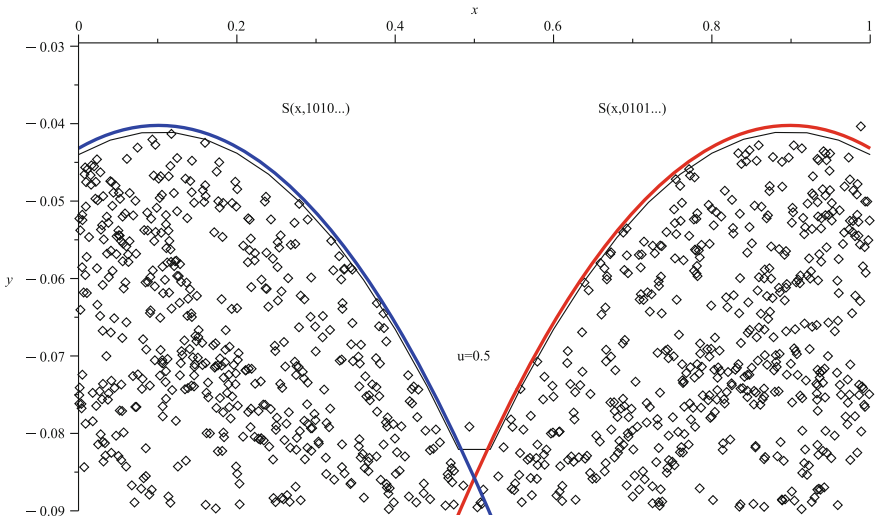


Fig. 14 The graph of $b(x)$ as an upper envelope agree with max $S(x, 1010...)$, $S(x, 01010...)$

In this way

$$S(0, (10)^\infty) = S(0, (01)^\infty) + \frac{6 + \lambda + 2}{4\,(4-\lambda)\,(2+\lambda)}.$$

Example 2 Here we consider $\lambda = 0.51$ and a periodic and continuous standard twist potential on the circle $A(x) = -(x - 0.5)^2$ that has the maximizing measure in a period two orbit, the superior envelope has two differentiable pieces since the unique turning point $u = 0.5$ pre-periodic. In the figure above, the dots are the iteration of F, the curves are $S(x, 10101...)$ and $S(x, 01010...)$ (Fig. 14). The curve dislocated is the graph of $b(x)$ computationally obtained as the superior envelope. The formal proof is given in the next.

Proposition 3 *Denote by* $b : S^1 \to \mathbb{R}$ *the function such that for* $0 \le x \le 1/2$, *we have* $b(x) = S(x, (10)^\infty)$, *and for* $1/2 \le x \le 1$, *we have* $b(x) = S(x, (01)^\infty)$. *Then,* b *is a* λ*-calibrated subaction for* $A(x) = -(x - 1/2)^2$, *that is, for any* $x \in S^1$, $b(x) = \max_i\{\lambda\, b(\tau_i x) + A(\tau_i x)\} =$

$$\max\{\lambda\, b(x/2) + A(x/2), \lambda\, b(x/2 + 1/2) + A(x/2 + 1/2)\}. \qquad (9)$$

Moreover, $b(0) = \frac{2\lambda}{4\,(4-\lambda)\,(2+\lambda)\,(\lambda-1)}$ *and this provides the explicit expression of* b.

Proof Note that for a given x we have

$$\lambda\, b(x/2) + A(x/2) = \lambda S(x/2, (10)^\infty) + (-1/4 + x/2 - x^2/4) =$$

$$\lambda\,[\,S(0, (10)^\infty) + \frac{(2\,\lambda - \lambda^2)\,x}{2\,(4-\lambda)\,(4-\lambda^2)} - \frac{1}{4(4-\lambda)}x^2] + (-1/4 + x/2 - x^2/4) =$$

$$\lambda S(0,(10)^\infty)+\lambda\frac{(2\lambda-\lambda^2)\,x}{2\,(4-\lambda)\,(4-\lambda^2)}-\lambda\frac{1}{4(4-\lambda)}x^2+(-1/4+x/2-x^2/4)=$$

$$S(0,(01)^\infty)\;-A(0)+\lambda\frac{(2\lambda-\lambda^2)\,x}{2\,(4-\lambda)\,(4-\lambda^2)}\;-$$

$$\lambda\frac{1}{4(4-\lambda)}x^2\;+(-1/4+x/2-x^2/4)=$$

$$S(0,(01)^\infty)\;+\lambda\frac{(2\lambda-\lambda^2)\,x}{2\,(4-\lambda)\,(4-\lambda^2)}-\lambda\frac{1}{4(4-\lambda)}x^2\;+x/2-x^2/4=$$

$$S(0,(01)^\infty)\;+\frac{(-\lambda^2-2\lambda+8)\,x}{(4-\lambda)\,(4-\lambda^2)}-\frac{x^2}{(4-\lambda)}\;=S(x,(01)^\infty),$$

As A is symmetric we claim that $S(x,(10)^\infty)=S((1-x),(01)^\infty)$.

Indeed,

$$S(x,(10)^\infty)=\sum_{k=0}\lambda^k A(\,\tau_{a_k}\circ\tau_{a_{k-1}}\circ\cdots\circ\tau_{a_0}\,(x)\,)=$$

$$A(\,\tau_1(x))+\lambda\,A(\,\tau_0\circ\tau_1(x))+\lambda^2\,A(\,\tau_1\circ\tau_0\circ\tau_1(x))+\cdots=$$

$$A(\,(x+1)/2)+\lambda\,A(\,\frac{1}{2}((x+1)/2))+\lambda^2\,A(\,\tau_1(\frac{1}{2}((x+1)/2))))+\cdots=$$

$$A(\,(x+1)/2)+\lambda\,A(\,\frac{1}{2}+(x/4\;-1/4))+\lambda^2\,A(\,\tau_1(\frac{1}{2}((x+1)/2))))+\cdots=$$

$$A(\,1/2-x)+\lambda\,A(\,\frac{1}{2}-(x/4\;-1/4))+\lambda^2\,A(\,\frac{(\frac{1}{2}((x+1)/2))+1}{2}))+\cdots=$$

$$A(\,\tau_0(1-x))+\lambda\,A(\,\tau_1\circ\tau_0(1-x))+\lambda^2\,A(\,\tau_0\circ\tau_1\circ\tau_0(1-x))+\cdots=$$

$$S((1-x),(01)^\infty).\tag{10}$$

Therefore, $b(x)=b(1-x)$. Moreover, $S(1/2,(10)^\infty)=S(1/2,(01)^\infty)$. Using this symmetry we get $\lambda\,b(x/2+1/2)+A(x/2+1/2)=S(x,(10)^\infty)$ from (10). From the above it follows (9). Note that from (8) we have

$$b(0)=\max\{\lambda\,b(0)+A(0),\lambda\,b(1/2)+A(1/2)\}=$$

$$\max\{\lambda\,b(0)-\;1/4,,\;\;\lambda\,b(1/2)\}=$$

$$\max\{\lambda\,S(0,(10)^\infty)-\;1/4,\;\;\lambda\,S(1/2,(10)^\infty)\}=$$

$$\max\{\lambda\, S(0,(10)^\infty) - 1/4,\ \lambda\, S(0,(10)^\infty) - \frac{2\,\lambda}{4\,(4-\lambda)\,(2+\lambda)}\} =$$

$$\max\{\lambda\, b(0) - 1/4,\ \lambda\, b(0) - \frac{2\,\lambda}{4\,(4-\lambda)\,(2+\lambda)}\} = \lambda\, b(0) - \frac{2\,\lambda}{4\,(4-\lambda)\,(2+\lambda)}.$$

In this way $S(0,(10)^\infty) = b(0) = \frac{2\,\lambda}{4\,(4-\lambda)\,(2+\lambda)\,(\lambda-1)}$.

8 Worked Examples and Computer Simulations

In the simulations we consider the function $S : (S^1, \{1,2,\ldots,d\}^{\mathbb{N}}) \to \mathbb{R}$ given by $S(x,a) = \sum_{k=0} \lambda^k A(\,(\tau_{a_k} \circ \tau_{a_{k-1}} \circ \cdots \circ \tau_{a_0})\,(x)\,)$, and, $a = (a_0, a_1, a_2, \ldots)$. The dynamics is defined by the inverse branches of $2x \bmod 1$, that is $\tau_0 = 0.5x$, $\tau_1 = 0.5x + 0.5$, $A(x)$ is a potential and $\lambda = 0.51$. We will build examples where $a = (a_0, a_1, a_2, \ldots)$ is truncated in a_7, and the dots represents the iteration of typical orbits by $F(x,s) = (T(x), \lambda\, s + A(x))$, $(x,s) \in S^1 \times \mathbb{R}$ producing a picture of the superior envelope of the attractor.

Example 3 Here we consider a periodic and continuous potential on the circle $A(x) = -(x-0.5)^2 + \varepsilon\psi(x) - drift$ for $\varepsilon = 0.05, drift = 0.2$ and $\psi(x) = (x - x^2)(1 + 3\,x + 9/2\,x^2 + 9/2\,x^3 + \frac{27}{8}\,x^4 + \frac{81}{40}\,x^5)$. Since, $-(x-0.5)^2$ is twist and has the maximizing measure in a period two orbits the same is true for A, but in this case, the superior envelope has three differentiable pieces and turning points $u = 0.21...$ and $v = 0.60....$ In the figure above, the dots are the iteration of F, the curves are $S(x, 11010...)$, $S(x, 10101...)$ and $S(x, 01010...)$ (Fig. 15). The curve dislocated is the graph of $b(x)$ computationally obtained as the superior envelope:

$$b(x) := \begin{cases} S(x, 110101....)\ ,\ 0 < x \le u \\ S(x, 101010....)\ ,\ u < x \le v \\ S(x, 010101....)\ ,\ v < x \le 1 \end{cases}.$$

Example 4 Here we consider a periodic and continuous potential on the circle

$$A(x) = \begin{cases} 6x - 3, & x < 1/2 \\ -6x + 3, & x \ge 1/2 \end{cases}$$

that is not twist but in this case, the superior envelope has two differentiable pieces since the unique turning point is pre-periodic according to Corollary 3. In the figure above, the dots are obtained by the iteration of F in an initial point and the curves are the graphs of $S(x, 10101...)$ and $S(x, 01010...)$ (Fig. 16). The curve slightly dislocated is the graph of $b(x)$ computationally obtained as the superior envelope.

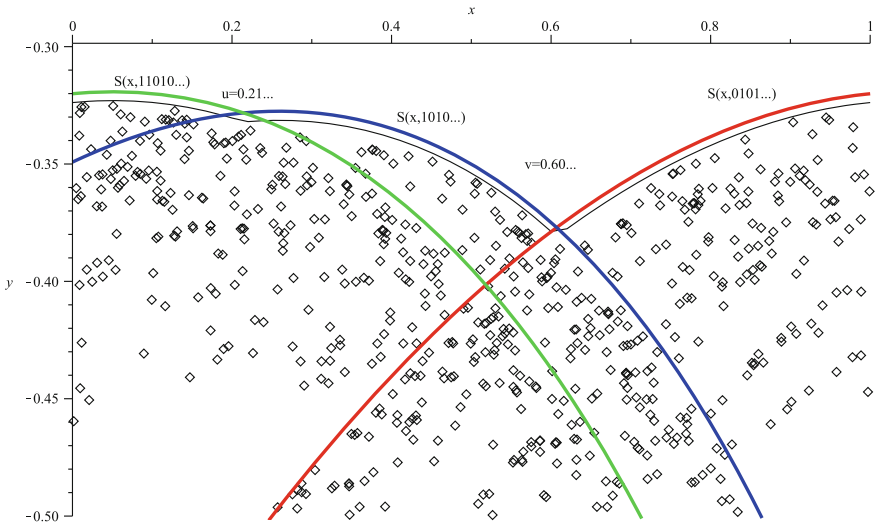


Fig. 15 Comparison between the upper envelope and b as union of graphs

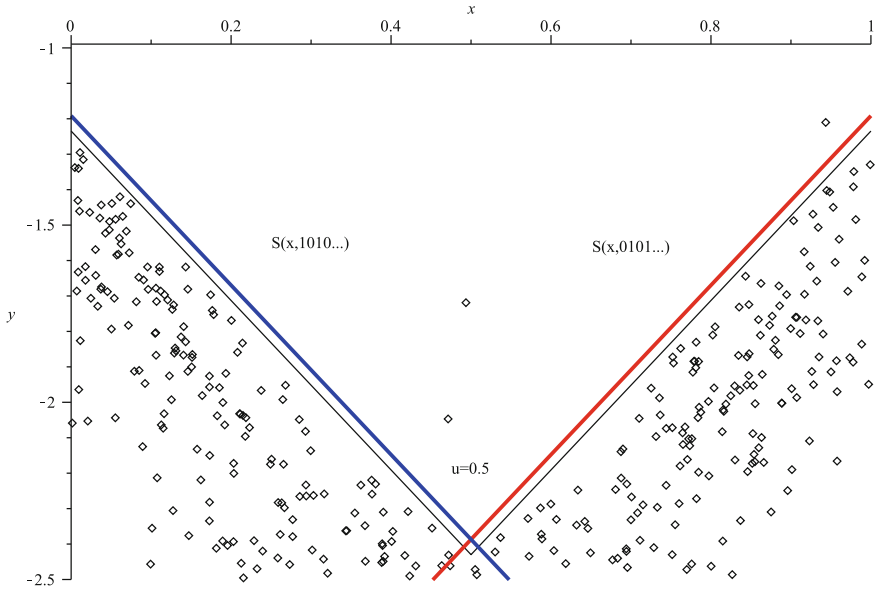


Fig. 16 Comparison between the upper envelope and b as union of graphs

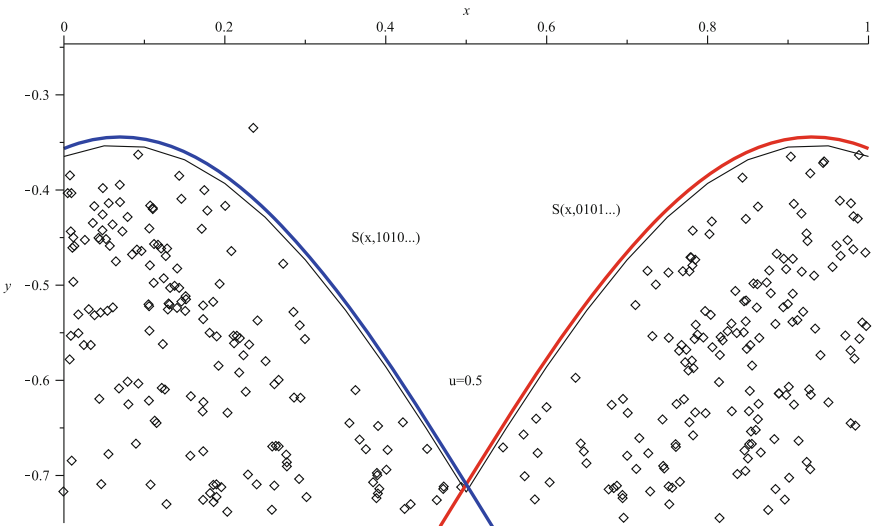


Fig. 17 Comparison between the upper envelope and b as union of graphs

Example 5 Here we consider a periodic and differentiable potential on the circle $A(x) = -1/2 - 1/2 \cos(2\pi x)$ that is not necessarily twist but in this case, the superior envelope has two differentiable pieces since the unique turning point is pre-periodic according to Corollary 3. In the figure above, the dots are the iteration of F in an initial point and the curves are defined by $S(x, 10101...)$ and $S(x, 01010...)$. The curve slightly dislocated is the graph of $b(x)$ computationally obtained as the superior envelope (Fig. 17).

9 Ergodic Transport

In this section A is assumed to be just Lipschitz. Following the notation of Sect. 2 we point out that: given $x = x_0$, there exists a sequence $x_k \in S^1$, $k \in \mathbb{N}$, such that $b(x_{k-1}) - \lambda b(\tau_{i_k}(x_k)) - A(\tau_{i_k}(x_k)) = 0$. One can consider the probability $m_n = \sum_{j=0}^{n-1} \frac{1}{n} \delta_{\sigma^j(a)}$, where σ is the shift, and, $a = a(x_0)$ is optimal for x_0. We define the probability μ_λ^* in $\{1, 2, \ldots, d\}^{\mathbb{N}}$, as any weak limit of a convergent subsequence m_{n_k}, $k \to \infty$ (which will be σ invariant).

Definition 12 We call μ_λ^* a λ-dual probability for A.

Note that from Proposition 1 if z is in the support of the λ-maximizing probability μ_λ, then $a(z)$ can be taken as periodic orbit for σ. In this case following the above reasoning we can produce a certain μ_λ^* which has support in a periodic orbit.

Consider a fixed $\overline{x} \in S^1$. Remember that we denote

$$A^*(a) = [\lambda S(\overline{x}, \sigma(a)) - S(\overline{x}, a)],$$

and in this way we get that for any (x, a) $A^*(a) = A(\tau_{a_0}(x)) + [\lambda W(\tau_{a_0}(x), \sigma(a)) - W(x, a)]$, where $W(x, a) = S(x, a) - S(\overline{x}, a)$. We called such W the λ-involution kernel for A. We called A^* is the λ-dual potential of A. The main strategy is to get results for A from properties of A^*. This is similar to the approach via primal and dual problems in Linear Programming. Note that W depends on the $\overline{x}$ we choose. Therefore, $A^* = A^*_{\overline{x}}$ depends of the $\overline{x}$. If we consider another base point x_1 instead $\overline{x}$, in order to get a different $W_1(x, a) = S(x, a) - S(x_1, a)$, then one can show that the corresponding A^*_1 (to A and W_1) satisfies $A^*_1 = A^* + \lambda\,(g \circ \sigma) - g$, for some continuous g. Note that $W - W_1$ just depends on a.

For the dual problem it will be necessary to consider the following problem: finding a function $b^* = b^*_\lambda$ which satisfies for all $a \in \Sigma$

$$\lambda b^*(a) = \max_{\sigma(c)=a} \{\, b^*(c) + A^*(c)\}.$$

In fact one can do more, it is possible to find a continuous function b^* that solves $\lambda b^*(\sigma(c)) = b^*(c) + A^*(c)$, $\forall c \in \Sigma$.

Just take, as in [3], $b^*(c) = -\sum_{j=0}^{\infty} \lambda^j\, A^*(\sigma^j(c)) = -\sum_{j=0}^{\infty} \lambda^j\, [\lambda S(\overline{x}, \sigma^{j+1}(c)) - S(\overline{x}, \sigma^j(c))] = -S(\overline{x}, c)$. In this case the corresponding rate function in the dual problem $R^*(c) = \lambda b^*(\sigma(c)) - b^*(c) - A^*(c)$ is constant equal zero. This situation is quite different from the analogous dual problem in [30].

Definition 13 We call b^*_λ **the dual λ-calibrated subaction.**

We assume, without lost of generality, that $A > 0$. Then, $b > 0$. It is natural to consider the sum $\sum R^*(\sigma^n)(z)$ in the dual problem (see [4, 13, 30]) but now this sum is zero. The role of the dual subactions V and V^* of [30] are now played by b and b^*, which are, respectively, the λ-calibrated subactions for A and A^*. Note that for all (x, a) $(b^* + b - W)(x, a) = -S(\overline{x}, a) + b(x) + S(\overline{x}, a) - S(x, a) = b(x) - S(x, a) \geq 0$. If a is a realizer for x, then $(b^* + b - W)(x, a) = 0$. Given A (and, a certain choice of A^* and W) the next result claims that the dual of R is R^* (which is constant equal zero), and the corresponding involution kernel is $(b^* + b - W)$.

Proposition 4

$$R(\tau_w x) = (b^* + b - W)(x, w) - \lambda(b^* + b - W)(\tau_w x, \sigma(w)).$$

Proof We know that $\lambda b^*(\sigma(w)) - b^*(w) = A^*(w)$, and, now using $x = T(\tau_w x)$, we get

$$b(x) - \lambda b(\tau_w x) = b(T(\tau_w x)) - \lambda b(\tau_w x) =$$

$$-A(\tau_w x) + A(\tau_w x) = R(\tau_w x) + A(\tau_w x).$$

Substituting the above in the previous equation we get

$$(b^* + b - W)(x, w) - \lambda\,(b^* + b - W)(\tau_w(x), \sigma(w)) =$$

$$[b^*(w) - \lambda\, b^*(\sigma(w))] + [b(x) - \lambda b(\tau_w x)] - W(x, w) + \lambda W(\tau_w x, \sigma(w)) =$$

$$-A^*(w) + R(\tau_w(x)) + A(\tau_w(x)) + \lambda\, W(\tau_w(x), \sigma(w)) - W(x, w) = R(\tau_w(x)),$$

because $A^*(w) = A(\tau_w x) + \lambda\, W(\tau_w x, \sigma(w)) - W(x, w)$. So the claim follows.

We present now a brief outline of Transport Theory (see [38, 39] as a general reference).

Definition 14 We denote by $\mathscr{K}(\mu, \mu^*)$ the set of probabilities $\hat{\eta}(x, w)$ on $\hat{\Sigma} = S^1 \times \Sigma$, such that $\pi_x^*(\hat{\eta}) = \mu$, and $\pi_w^*(\hat{\eta}) = \mu^*$. Each element in $\mathscr{K}(\mu, \mu^*)$ is called a plan.

In Transport Theory one is interested in plans which minimize the integral a given lower semi-continuous cost $c : \Sigma \to \mathbb{R}$. The Classical Transport Theory is not a Dynamical Theory. It is necessary to consider a dynamically defined cost in order to be able to get some results such that the optimal plan is invariant for some dynamics. We are going to consider below the cost function $c(x, w) = -W(x, w) = -W_\lambda(x, w)$ where $W(x, w)$ is a λ-involution kernel of the Lipschitz potential A. The Kantorovich Transport Problem: consider the minimization problem

$$C(\mu, \mu^*) = \inf_{\hat{\eta} \in \mathscr{K}(\mu, \mu^*)} \int \int -W(x, w)\, d\,\hat{\eta}.$$

Definition 15 A probability $\hat{\eta}$ on $\hat{\Sigma}$ which attains such infimum is called an optimal transport probability, or, an optimal plan, for $c = -W$.

It is natural to consider the bijective transformation $\mathbb{T}$ which acts on $\hat{\Sigma} = S^1 \times \Sigma$ in such way that $\mathbb{T}^{-1}(x, w) = (\tau_w x, \sigma(w))$. We will show later that for μ_λ and μ_λ^* there exists a $\mathbb{T}$-invariant probability $\hat{\mu}_{min}$ which attains the optimal transport cost. Dynamically defined costs can determine optimal plans which have dynamical properties.

Definition 16 A pair of continuous functions $f(x)$ and $f^\#(w)$ will be called c-admissible (or, just admissible for short) if

$$f^\#(w) = \min_{x \in S^1} \{-f(x) + c(x, w)\}.$$

We denote by $\mathscr{F}$ the set of admissible pairs. The Kantorovich dual Problem: given the cost $c(x, w)$ consider the maximization problem

$$D(\mu, \mu^*) = \max_{(f, f^\#) \in \mathscr{F}} \left(\int f d\mu + \int f^\# d\mu^* \right).$$

In this problem one is interested in any pair (when exists) $(f, f^{\#}) \in \mathscr{F}$ which realizes the maximum in the right side of the above expression.

Definition 17 A pair of admissible $(f, f^{\#}) \in \mathscr{F}$ which attains the maximum value will be called an optimal Kantorovich pair.

Under quite general conditions [38] (which are satisfied here) $D(\mu, \mu^*) = C(\mu, \mu^*)$. We denote $\Gamma = \Gamma_b = \{(x, w) \in S^1 \times \Sigma \mid b(x) = (-b^* + W)(x, w)\}$. A classical result in Transport Theory [38]: if $\hat{\eta}$ is a probability in $\mathscr{K}(\mu, \mu^*)$, $(f, f^{\#})$ is an admissible pair, and the support of $\hat{\eta}$ is contained in the set $\{(x, w) \in \hat{\Sigma} \mid \text{such that } (f(x) + f^{\#}(w)) = c(x, w)\}$, then, $\hat{\eta}$ in an optimal plan for c and $(f, f^{\#})$ is an optimal pair in $\mathscr{F}$.

This is the so called slackness condition of Linear Programming (see [39] Remark 5.13 p. 59). This results allows one to get in some cases the solution of the primal problem (which is looking for optimal plans) via de dual problem (which is looking for optimal pairs of functions). If you have a good guess that a certain $\hat{\eta}$ is the optimal plan you can try to find an admissible pair satisfying the above condition on the support of the plan. If you succeeded then you show that the plan $\hat{\eta}$ is indeed the solution of the transport problem. This is the power of the dual problem approach.

We will show that for the problem $D(\mu_\lambda, \mu_\lambda^*)$ the functions $-b$ and $-b^*$ define an optimal Kantorovich pair. From this fact becomes clear the importance of the set Γ.

Our main result in this section is:

Theorem 6 *For the probabilities μ_λ, μ_λ^* and the cost $-W$, the associated transport problem is such that the functions $-b$ and $-b^*$ define an optimal Kantorovich pair, and, the optimal plan is invariant by $\mathbb{T}$.*

Proof We claim first that $-b$ and $-b^*$ are $-W$-admissible. Indeed, $p(x, w) := (b^* + b - W)(x, w) \geq 0$. Moreover, for each x there exists a w which is a realizer and then $p(x, w) = 0$. Therefore, for each x we have that

$$b(x) = \max_{w \in \Sigma}\{-b^*(w) + W(x, w)\} = \max_{w \in \Sigma} S(x, w). \tag{11}$$

For each x we denote $w_x \in \Sigma$ the realizer for the above equation. We can say that b is the W transform of $-b^*$ [38, 39]. Note that

$$\Gamma = \{(x, w) \in S^1 \times \Sigma \mid p(x, w) = 0\}.$$

We will show that the infimum of the cost $-W$, denoted $c(A, \lambda)$, is equal to $\int -b^* \, d\mu_\lambda^* + \int -b \, d\mu_\lambda$.

The next proposition is similar to a result on [30]. Remember that $R = -(A - b \circ T + \lambda b) \geq 0$ is called the rate function.

Proposition 5 (Fundamental relation) *For any* (x, w)

$$R(\tau_w x) = p(x, w) - \lambda \, p(\tau_w x, \sigma(w)) \tag{12}$$

Moreover, if $\mathbb{T}^{-1}(x, w) = (\tau_w x, \sigma(w))$, *then*

(a) $p - \lambda\, p \circ \mathbb{T}^{-1}(x, w) = R(\tau_w x) \geq 0$;
(b) Γ *is invariant by the action of* $\mathbb{T}^{-1}$;
(c) *if* $a = (i_0, i_1, i_2, ...)$ *is optimal for* x, *then* $\sigma^n(a)$ *is optimal for* $(\tau_{i_{n-1}} \circ \cdots \circ \tau_{i_1} \circ \tau_{i_0})(x)$.

Proof The first claim (a) is a trivial consequence of the definition of $\mathbb{T}^{-1}$. The second one it is a consequence of: $p \geq 0$, and

$$p - \lambda\,(p \circ \mathbb{T}^{-1})(x, w) \geq 0 \Rightarrow p(x, w) \geq \lambda\,(p \circ \mathbb{T}^{-1})(x, w).$$

From the above we get that in the case (x, w) is optimal, then, $\mathbb{T}^{-1}(x, w)$ is also optimal. Indeed, we have that $p(x, w) = 0 \rightarrow p(\tau_w(x), \sigma(w)) = 0$. Item (c) follows by induction.

In this way $\mathbb{T}^{-n}$ **spread optimal pairs**. This is a nice property that has no counterpart in the Classical Transport Theory.

Take now $(z_0, w_0) \in \Gamma_V$ and, for each n, $\hat{\mu}_n = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\mathbb{T}^{-j}(z_0, w_0)}$. Note that $\mathbb{T}^{-j}(z_0, w_0)$ is optimal. The closure of the set $\{\mathbb{T}^{-j}(z_0, w_0),\ j \in \mathbb{N}\}$ is contained in the support of the optimal transport plan.

Proposition 6 *We claim that any weak limit of convergent subsequence* $\hat{\mu}_{n_k}$, $k \to \infty$, *will define a probability* $\hat{\mu}$ *which is optimal for the transport problem for* $-W$ *and its marginals. In this way we will show the existence of a* $\mathbb{T}$*-invariant probability on* $S^1 \times \Sigma$ *which is optimal for the associated transport problem.*

Proof Indeed, we considered before a certain z_0, its realizer w_0, and then a convergent subsequence μ_{n_k} (notation of last section), $n_k \to \infty$, in order to get μ_λ. If we consider above the corresponding subsequence $\mathbb{T}^{-n_k}(z_0, w_0)$ we get that the projection of $\hat{\mu}$ on the S^1 coordinate is μ_λ.

In an analogous way, we consider as before a certain z_0, its realizer w_0, and then a convergent subsequence m_k to define μ^*_λ. If we consider above a subsequence m_k of the previous sequence n_k (last paragraph) we get that the projection of $\hat{\mu}$ on the Σ coordinate is μ^*_λ. As $p(x, w) = (b^* + b - W)(x, w)$ and p is zero on the orbit $\mathbb{T}^{-n_k}(z_0, w_0)$ we get that g is also zero in the support of any associated weak convergent subsequence. Then any probability $\hat{\mu}$ obtained in this way is such that projects respectively on μ_λ and μ^*_λ, and, moreover, satisfies

$$\int -W d\hat{\mu} = \int (-b^*) d\mu^* + \int (-b)\, d\mu_\lambda.$$

Therefore, $C(\mu, \mu^*) = \int (-b^*) d\mu^* + \int (-b) d\mu_\lambda$.

We point out that for the purpose of proving the conjecture the next proposition is the key result. It is just a trivial consequence of Theorem 6 and expression (11).

Proposition 7 *Suppose that A is Lipschitz, the maximizing probability μ_λ has support in a unique periodic orbit of period k and μ_λ^* is a dual λ-maximizer with support on the dual periodic orbit of period k for σ, then*

$$b(x) = \max_{w\in\Sigma}\{-b^*(w) + W(x,w)\} = -b^*(a) + W(x,a) \; =$$

$S(x,a) = \max_{w\in\Sigma} S(x,w)$, *where $a = a(x)$ is the periodic realizer of x. In this case a is in the support of μ_λ^*. Moreover, the procedure: given $(z_0, a(z_0)) \in \Gamma_V$ take $\hat{\nu}$*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\mathbb{T}^{-j}(z_0,a(z_0))} \; = \; \hat{\nu},$$

is such that $\hat{\nu}$ is optimal and has support on a periodic orbit for $\mathbb{T}$. In the support of $\hat{\nu}$ we have $b(x) + b^*(a(x)) = W(x, a(x))$.

Proposition 8 *Suppose W satisfies a twist condition. Denote by $\mathfrak{w} : S^1 \to \Sigma$ the function such that for a given x we have that $\mathfrak{w}(x)$ is a choice of the eventual possible w_x as defined above. Then, $\mathfrak{w}$ is monotonous non-decreasing (using the lexicographic order in Σ).*

The proof of this proposition is the same as the one in Proposition 6.2 in [30] or Proposition 2.1 in [13]. In [28] other kinds of results in Ergodic Transport Theory are considered.

Consider $0 < \lambda < 1$, and the map $G(w,s) = (\sigma(w), \lambda\, s + A^*(w))$, where $G : \{1,2,\ldots,d\}^{\mathbb{N}} \times \mathbb{R} \to \{1,2,\ldots,d\}^{\mathbb{N}} \times \mathbb{R}$, and $A^* : \{1,2,\ldots,d\}^{\mathbb{N}} \to \mathbb{R}$ is the dual potential.

The dynamics of attractor for F has associated to it a dual repellor naturally defined by G acting on $\{1,2,\ldots,d\}^{\mathbb{N}} \times \mathbb{R}$. The boundary of the repellor set is the graph of $b^* : \{1,2,\ldots,d\}^{\mathbb{N}} \to \mathbb{R}$ which is the λ-dual calibrated subaction.

References

1. Alexander, J., Yorke, J.: Fat bakers transformations. Ergod. Theory Dyn. Syst. **4**, 1–23 (1984)
2. Avila, A., Gouzel, S., Tsujii, M.: M. Smoothness of solenoidal attractors. Discret Contin. Dyn. Syst. **15**(1), 21–35 (2006)
3. Baladi, V., Smania, D.: Smooth deformations of piecewise expanding unimodal maps. Discret. Contin. Dyn. Syst. **23**(3), 685–703 (2009)
4. Baraviera, A., Lopes, A.O., Thieullen, P.: A large deviation principle for equilibrium states of Hölder potencials: the zero temperature case. Stoch. Dyn. **6**, 77–96 (2006)
5. Baraviera, A.T., Cioletti, L.M., Lopes, A.O., Mohr, J., Souza, R.R.: On the general one-dimensional XY model: positive and zero temperature, selection and non-selection. Rev. Math. Phys. **23**(10), 1063–1113 (2011)
6. Baraviera, A., Leplaideur, R., Lopes, A. O. : Ergodic Optimization, zero temperature limits and the max-plus algebra. In: mini-course in XXIX Colóquio Brasileiro de Matemática - IMPA - Rio de Janeiro (2013)

7. Bamón, R., Kiwi, J., Rivera-Letelier, J., Urzúa, R.: On the topology of solenoidal attractors of the cylinder. Ann. Inst. H. Poincaré Anal. Non Lineaire **23**(2), 209–236 (2006)
8. Bhattacharya, P., Majumdar, M.: Random Dynamical Systems. Cambridge University Press, Cambridge (2007)
9. Bousch, T.: Le poisson n'a pas d'arêtes. Ann. Inst. H. Poincaré, Probab. Stat. **36**(4), 489–508 (2000)
10. Bousch, T.: La condition de Walters. Ann. Sci. ENS **34**, 287–311 (2001)
11. Contreras, G.: Ground states are generically a periodic orbit. Invent. Math. **205**(2), 383–412 (2016)
12. Contreras, G., Lopes, A.O., Thieullen, Ph: Lyapunov minimizing measures for expanding maps of the circle. Ergod. Theory Dyn. Syst. **21**, 1379–1409 (2001)
13. Contreras, G., Lopes, A.O., Oliveira, E.R.: Ergodic Transport Theory, periodic maximizing probabilities and the twist condition. In: Zilberman, D., Pinto, A., Modeling, Optimization, Dynamics and Bioeconomy, Springer Proceedings in Mathematics, pp. 183–219 (2014)
14. Conze, J.P., Guivarc'h, Y.: Croissance des sommes ergodiques et principe variationnel, manuscript, circa 1993
15. Delon, J., Salomon, J., Sobolevski, A.: Fast transport optimization for Monge costs on the circle. SIAM J. Appl. Math **70**(7), 2239–2258 (2010)
16. Garibaldi, E., Lopes, A.O.: On Aubry-Mather theory for symbolic dynamics. Ergod. Theory Dyn. Syst. **28**(3), 791–815 (2008)
17. Gomes, D.A.: Viscosity Solution methods and discrete Aubry-Mather problem. Discret. Contin. Dyn. Syst. **13**(1), 103–116 (2005)
18. Gomes, D.A.: Generalized Mather problem and selection principles for viscosity solutions and Mather measures. Adv. Calc. Var. **1**, 291–307 (2008)
19. Grebogi, C., Nusse, H., Ott, E., Yorke, J.: Basic sets: sets that determine the dimension of basin boundaries. Dynamical Systems (College Park, MD, 1986–250. Lecture Notes in Mathematics, vol. 1342. Springer, Berlin (1988)
20. He, B., Gan, S.: Robustly non-hyperbolic transitive endomorphisms on $\mathbb{T}^2$. Proc. Am. Math. Soc. **141**, 2453–2465 (2013)
21. Iglesias, J., Portela, A., Rovella, A., Xavier, J.: Attracting sets on surfaces. Proc. Am. Math. Soc. **143**(2), 765–779 (2015)
22. Iturriaga, R., Sanchez-Morgado, H.: Limit of the infinite horizon discounted Hamilton-Jacobi equation. Discret. Contin. Dyn. Syst. Ser. B **15**(3), 623–635 (2011)
23. Iturraiga, R., Lopes, A., Mengue, J.: Selection of calibrated subaction when temperature goes to zero in the discounted problem, preprint UFRGS
24. Jenkinson, O.: Ergodic optimization. Discret. Contin. Dyn. Syst. Ser. A **15**, 197–224 (2006)
25. Jenkinson, O.: Optimization and majorization of invariant measures. Electron. Res. Announc. Am. Math. Soc. **13**, 1–12 (2007)
26. Jenkinson, O.: A partial order on x2-invariant measures. Math. Res. Lett. **15**(5), 893–900 (2008)
27. Lopes, A.O.: Thermodynamic Formalism, Maximizing Probabilities and Large Deviations. Lecture Notes, Dynamique en Cornouaille, France (2011). http://mat.ufrgs.br/~alopes/hom/notesformtherm.pdf
28. Lopes, A.O., Menge, J.K.: Duality theorems in ergodic transport. J. Stat. Phys. **149**(5), 921–942 (2012)
29. Lopes, A.O., Mohr, J.: Semiclassical limits, Lagrangian states and coboundary equations. Stoch. Dyn. **17**(2), 1750014 (19 pages) (2017)
30. Lopes, A.O., Oliveira, E.R., Smania, D.: Ergodic transport theory and piecewise analytic subactions for analytic dynamics. Bull. Braz. Math. Soc. **43**(3), 467–512 (2012)
31. Lopes, A.O., Oliveira, E.R., Thieullen, P.: The dual potential, the involution kernel and transport in ergodic optimization. In: Bourguignon, J.-P., Jelstch, R., Pinto, A., Viana, M. (eds.) Dynamics, Games and Science -International Conference and Advanced School Planet Earth DGS II, Portugal (2013), pp 357–398. Springer, Berlin (2015)
32. Lopes, A.O., Mengue, J.K., Mohr, J., Souza, R.R.: Entropy and variational principle for one-dimensional lattice systems with a general a-priori probability: positive and zero temperature. Ergod. Theory Dyn. Syst. **35**(6), 1925–1961 (2015)

33. Lopes, A.O., Mengue, J.K., Mohr, J., Souza, R R.: Entropy, pressure and duality for Gibbs plans in ergodic transport. Bull. Braz. Math. Soc. **46**(3), 353–389 (2015)
34. Park, B.-S., Grebogi, C., Ott, E., Yorke, J.: Scaling of fractal basin boundaries near intermittency transitions to chaos. Phys. Rev. A **40**(3), 1576–1581 (1989)
35. Robinson, C.: Dynamical Systems. CRC Press, London (1995)
36. Souza, R.R.: Ergodic and thermodynamic games. Stoch. Dyn. **16**(2), 1660008-1-15 (2016)
37. Tsujii, M.: Fat solenoidal attractors. Nonlinearity **14**, 1011–1027 (2001)
38. Villani, C.: Topics in Optimal Transportation. AMS, Providence (2003)
39. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2009)

Transport and Large Deviations for Schrodinger Operators and Mather Measures

A. O. Lopes and Ph. Thieullen

Abstract In this mainly survey paper we consider the Lagrangian $L(x, v) = \frac{1}{2}|v|^2 - V(x)$, and a closed form w on the torus $\mathbb{T}^n$. For the associated Hamiltonian we consider the the Schrodinger operator $\mathbf{H}_\beta = -\frac{1}{2\beta^2}\Delta + V$ where β is large real parameter. Moreover, for the given form $\beta\, w$ we consider the associated twist operator $\mathbf{H}^w_\beta$. We denote by $(\mathbf{H}^w_\beta)^*$ the corresponding backward operator. We are interested in the positive eigenfunction ψ_β associated to the the eigenvalue E_β for the operator $\mathbf{H}^w_\beta$. We denote ψ^*_β the positive eigenfunction associated to the eigenvalue E_β for the operator $(\mathbf{H}^w_\beta)^*$. Finally, we analyze the asymptotic limit of the probability $\nu_\beta = \psi_\beta\, \psi^*_\beta$ on the torus when $\beta \to \infty$. The limit probability is a Mather measure. We consider Large deviations properties and we derive a result on Transport Theory. We denote $L^-(x, v) = \frac{1}{2}|v|^2 - V(x) - w_x(v)$ and $L^+(x, v) = \frac{1}{2}|v|^2 - V(x) + w_x(v)$. We are interest in the transport problem from μ_- (the Mather measure for L^-) to μ_+ (the Mather measure for L^+) for some natural cost function. In the case the maximizing probability is unique we use a Large Deviation Principle due to N. Anantharaman in order to show that the conjugated sub-solutions u and u^* define an admissible pair which is optimal for the dual Kantorovich problem.

Keywords Schrodinger operators · Eigenfunction and eigenvalue of the Hamiltonian operator · Transport · Involution Kernel · Large Deviations Mather measure · Viscosity solutions

A. O. Lopes (✉)
Instituto de Matemática-UFRGS, Avenida Bento Gonçalves,
Porto Alegre-RS 9500, Brazil
e-mail: alopes@mat.ufrgs.br

Ph. Thieullen
Institut de Mathématiques, Université Bordeaux, F-33405 Talence, France
e-mail: philippe.thieullen@math.u-bordeaux1.fr

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_11

1 Introduction and Basic Definitions

Given a closed form w on the torus $\mathbb{T}^n$ we consider the Lagrangian $L(x,v) = \frac{1}{2}|v|^2 - V(x) + w$, where $L : T\mathbb{T}^n \to \mathbb{R}$ and $T\mathbb{T}^n$ is the tangent bundle.

The infimum of $\int L(x,v)\, d\mu(x,v)$ among the invariant probabilities for the Euler Lagrange flow on the tangent bundle $T\mathbb{T}^n$ is called the critical value of L. A probability which attains such infimum is called a Mather measure (see [5] for references an general results).

We denote by $H(x,p) = \frac{1}{2}|p|^2 + V(x)$ the associated Hamiltonian for the Lagrangian $L(x,v) = \frac{1}{2}|v|^2 - V(x)$ and for each $\beta \in \mathbb{R}$ we consider the corresponding Schrodinger operator $\mathbf{H}_\beta = -\frac{1}{2\beta^2}\Delta + V$ for such Hamiltonian.

For each β we consider a certain associated quantum state and quantum probability on $\mathscr{L}^2(\mathbb{T}^n)$ (associated to an eigenvalue of $\mathbf{H}_\beta$) and we are interested in the limit of such probability when $\beta \to \infty$.

We call β the semiclassical parameter. In an alternative form we can take $\hbar = \frac{1}{\beta}$ and consider the limit when $\hbar \to 0$.

An interesting relation of such limit probabilities with Mather measures was investigated by N. Anantharaman (see [1–3]).

We will present here some of these results which are related to transport and large deviation properties.

Consider $w(v) =< P, v >$ a closed form w in the torus $\mathbb{T}^n$, where P is a vector in $\mathbb{R}^n$.

Suppose that μ_+ and μ_- are respectively the Mather measures for the Lagrangians

$$L^+(x,v) = \frac{1}{2}|v|^2 - V(x) + w(v) \text{ and } L^-(x,v) = \frac{1}{2}|v|^2 - V(x) - w(v),$$

$x \in \mathbb{T}^n$ and $V : \mathbb{T}^n \to \mathbb{R}$ smooth.

We assume the Mather measure is unique in each problem (see [5, 10, 11]).

We will follow closely the notation of the nice exposition [1] (see also [2, 3]). The results presented here in the future sections are inspired in [14]. The main tool is the involution kernel introduced in [4] (see also [6, 12, 15–19]).

We will consider the Lax-Oleinik operator T_t^-, $t \geq 0$, given by

$$T_t^- u_1(x) = \inf_{\gamma(t)=x,\, \gamma:[0,t]\to\mathbb{T}^n} \{u_1(\gamma(0)) + \int_0^t L^-(\gamma(s), \gamma'(s))ds\}.$$

Denote by u (Lipchitz), $u : \mathbb{T}^n \to \mathbb{R}$, the unique (up to additive constant because μ_- is unique) solution of $T_t^- u = u + t\, E$, for all $t \geq 0$, and where E is constant.

Consider the Lax-Oleinik operator T_t^+, $t \geq 0$, given by

$$T_t^+ u_2(x) = \inf_{\gamma(0)=x,\, \gamma:[0,t]\to\mathbb{T}^n} \{u_2(\gamma(t)) - \int_0^t L^-(\gamma(s), \gamma'(s))ds\}.$$

Denote by u^* the Lipchitz function $u^* : \mathbb{T}^n \to \mathbb{R}$, such that, $T_t^+(-u^*) = -u^* + E\,t$.

We assume that u and u^* are such $u + u^*$ is zero in the support of μ_-.

The function $W(x, y)$ (which could be called the convolution kernel) is given by the below expression

$$-\inf_{\alpha \in C^1([0,1],\mathbb{T}^n),\, \alpha(0)=x,\, \alpha(1)=y} \{\int_0^1 [-V(\alpha(s)) + w(\alpha'(s))]\, ds + \int_0^1 \frac{1}{2}\, ||\alpha'(t)||^2\}.$$

We denote by $h(y, y)$ the Peierls barrier for the Lagrangian L^-. In the present case $h(y, y) = u(y) + u^*(y)$.

Main references on Transport Theory are [20–22].

We denote by $\mathscr{K}(\mu_+, \mu_-)$ the set of probabilities $\hat{\mu}$ on $\mathbb{T}^n \times \mathbb{T}^n$, such that respectively $\mu_+ = \pi_1^{\#}(\hat{\mu})$ and $\mu_- = \pi_2^{\#}(\hat{\mu})$.

Give $c(x, y)$ we say that f and g are c-admissible if, for any $x, y \in \mathbb{R}^n$, we have $f(x) - g(y) \leq c(x, y)$. We denote by $\mathscr{F}$ the set of such pairs (f, g).

We will consider, for the cost function $c(x, y) = -W(y, x)$, a c-Kantorovich problem

$$\inf_{\hat{\mu} \in \mathscr{K}(\mu_+,\mu_-)} \int\int c(x, y)\, d\hat{\mu}(x, y).$$

We denote the minimizing probability by $\hat{\mu}_{\min}$. Note that this probability projects on the second variable on μ_-.

Note that the transport optimal probability for $-W$ and for $-W + I$ (where I is the Peierl's barrier) are the same.

We point out that the projected Mather measures μ^+ and μ^- are the same in the present case.

We will show here that the dual problem for $-W$

$$\max\{\int f(x)\, d\mu_+(x) - \int g(y)\, d\mu_-(y) \mid f(x) - g(y) \leq c(x, y)\} =$$

$$\max\{\int f(x)\, d\mu_+(x) - \int g(y)\, d\mu_-(y) \mid (f, g) \in \mathscr{F}\},$$

has a pair of optimal solutions (u, u^*) which are the viscosity solutions of the Hamilton-Jacobi equations (fixed points of the corresponding Lax-Oleinik operators as defined above)

We can consider alternatively (the same problem)

$$\inf_{\hat{\mu} \in \mathscr{K}(\mu_+,\mu_-)} \int\int \tilde{c}(x, y)\, d\hat{\mu}(x, y),$$

where $\tilde{c}(x, y) = -W(y, x) + h(y, y)$. The introduction of a function on the variable y which vanishes in the support of μ_- does not change the minimizing measure. However, this new problem have a different optimal pair.

We denote by $\mathscr{W}_x^h$ the Brownian motion in $\mathbb{R}^n$ (with coefficient h, that is, at time $t = 1$ the variance is $\sqrt{h}$) beginning at x, and $\mathscr{W}_{x,y,t}^h$ its disintegration at the point y and at the time t.

Consider the Schrodinger $\mathbf{H}^h = -\frac{h^2}{2}\Delta + V$ (where V is the periodic extension to $\mathbb{R}^n$) which acts on real (periodic) functions defined in $\mathbb{R}^n$. It is known that $\mathbf{H}$ has pure point spectrum (see [7, 13]).

Note that

$$-\frac{1}{h}\mathbf{H}^h = \frac{h}{2}\Delta - \frac{1}{h}V.$$

The Kernel $K(x, y, t)$ of the extension of $e^{-\frac{t}{h}H}$ to an integral operator is (see [1])

$$K(x, y, t) = \int e^{-\frac{1}{h}\int_0^t V(\alpha(s))\,ds}\, \mathscr{W}_{x,y,t}^h(d\alpha).$$

Given

$$L^w(x, v) = \frac{1}{2}v^2 - V(x) - w(v) = \frac{1}{2}v^2 - V(x) - < P, v >,$$

the corresponding Hamiltonian $H^w(x, p)$ via Legendre transform is

$$H^w(x, p) = \frac{||p + P||^2}{2} + V(x).$$

In the same way, for

$$L^+(x, v) = \frac{1}{2}v^2 - V(x) + w(v) = \frac{1}{2}v^2 - V(x) - < P, v >,$$

the corresponding Hamiltonian $H^{w^*}(x, p)$ is

$$H^{w^*}(x, p) = \frac{||p - P||^2}{2} + V(x).$$

Consider, a certain point $x_0 = \mathscr{O} \in \mathbb{R}^n$ fixed (on the universal cover of the torus). As the form w on the torus is closed, it is exact on the lifting to the universal cover, then, the value $\int_{x_0}^x w$ does not depend on the path we choose to connect x_0 to x.

2 Transport in the Configuration Space for the Aubry-Mather Problem

For each real value β we consider the operator

$$\mathbf{H}_\beta^w = e^{-\beta \int_{x_0}^x w} \circ \mathbf{H}_\beta \circ e^{\beta \int_{x_0}^x w} = e^{-\beta \int_{x_0}^x w} \circ \left(-\frac{1}{2\beta^2} \Delta + V \right) \circ e^{\beta \int_{x_0}^x w}.$$

We can consider such operator acting on the torus or on $\mathbb{R}^n$. When we consider the Brownian motion we should consider, off course, its action on $\mathbb{R}^n$.

The Kernel $K(x, y, t)$ of the extension of $e^{t\beta \mathbf{H}_\beta^w}$ to an integral operator is given by

$$K_\beta(x, y, t) = \int e^{-\beta \int_0^t V(\alpha(s))\, ds + \beta < P, (y-x) >} \mathscr{W}_{x,y,t}^{\beta^{-1}}(d\alpha).$$

Note that above we consider the integral

$$\beta \int_0^t [-V(\alpha(s)) + w_{\alpha(s)}(\alpha'(s))]\, ds.$$

$\mathbf{H}_\beta^w$ is not self adjoint but has a real pure point spectrum.

We denote by E_β the maximum eigenvalue of $\mathbf{H}_\beta^w$ (acting on real functions) and ψ_β is the corresponding normalized real eigenfunction in $\mathscr{L}^2(\mathbb{T}^n, dx)$. The positive eigenfunction ψ_β is unique if we assume its norm is 1. It's the only totally positive eigenfunction of $\mathbf{H}_\beta^w$ (see [1] expression (3.15)). The eigenvalue is simple and isolated (see Appendix on [2]).

For each real value β we consider the w-backward operator

$$\mathbf{H}_\beta^{w^*} = e^{\beta \int_{x_0}^x w} \circ \mathbf{H}_\beta \circ e^{-\beta \int_{x_0}^x w} = e^{\beta \int_{x_0}^x w} \circ \left(-\frac{1}{2\beta^2} \Delta + V \right) \circ e^{-\beta \int_{x_0}^x w}.$$

We will be interested in high values of β.

E_β is the maximum eigenvalue of $\mathbf{H}_\beta^{w^*}$ and we denote by ψ_β^* the corresponding real eigenfunction in $\mathscr{L}^2(\mathbb{T}^n, dx)$. Similar properties to the case of ψ_β are true for such eigenfunction. We assume $\int \psi_\beta^*(x)\, dx = 1$ and also $\int \psi_\beta(x)\, \psi_\beta^*(x)\, dx = 1$.

$\mathbf{H}_\beta^{w^*} \circ \mathbf{H}_\beta^w$ is self adjoint.

We will be interested here in the probabilities

$$\nu_\beta(dx) = \psi_\beta(x)\, \psi_\beta^*(x)\, dx.$$

The probability $\nu_\beta(dx) = \psi_\beta(x)\, \psi_\beta^*(x)\, dx$ is stationary for the Markov operator

$$Q^t(f)(x) = e^{-t E_w}\, \psi_\beta(x)^{-1} e^{t \mathbf{H}_\beta^w}(\psi_\beta\, f)(x)$$

on the torus $\mathbb{T}^n$ (see [2]).

The correct point of view is to consider ψ_β as an eigenfunction and $\rho_\beta = \psi_\beta^*(x)\,dx$ as an eigen-probability for the semi-group $t \to e^{t\,\mathbf{H}_\beta^w}$.

Consider

$$u_\beta = -\frac{\log \psi_\beta}{\beta} \quad \text{and} \quad u_\beta^* = -\frac{\log \psi_\beta^*}{\beta}.$$

It is known that the following equalities are true:

$$-\frac{1}{2\,\beta}\,\Delta\, u_\beta \,+\, H^w(x, d_x u_\beta) = E_\beta,$$

and

$$-\frac{1}{2\,\beta}\,\Delta\, u_\beta^* \,+\, H^w(x, -d_x u_\beta^*) = E_\beta,$$

The β-families of functions u_β and u_β^* are equi-Lipschizians and we can obtain from this fact convergent subsequences. We assume here the Mather measure is unique, and therefore the limits exist in the uniform convergence topology, that is

$$\lim_{\beta\to\infty} u_\beta = u \quad \text{and} \quad \lim_{\beta\to\infty} u_\beta^* = u^*.$$

It is known that $\lim_{\beta\to\infty} E_\beta$ exist and we denote this value by E.

By stability of the viscosity solutions, the limits u and u^* are, respectively, viscosity solutions of the equations

$$H^w(x, d_x\, u) = E \quad \text{and} \quad H^w(x, -d_x\, u^*) = E$$

We assume also that the Mather measure μ for the lagrangian L^w is unique. In this case it is known (see for instance [1, 2]) that in the weak topology

$$\lim_{\beta\to\infty} \nu_\beta = \mu.$$

In Proposition 3.11 in [1] the following Large Deviation Principle is obtained (see also [2, 3]):

Proposition 1 *Suppose the Mather measure is unique. Suppose also that in the uniform convergence topology*

$$\lim_{\beta\to\infty} u_\beta = u \quad \textit{and} \quad \lim_{\beta\to\infty} u_\beta^* = u^*.$$

Then, for $I(x) = u(x) + u^*(x)$ (from the normalization we choose before $I(x) \geq 0$), we have

(1) for any open set $O \subset \mathbb{T}^n$,

$$\liminf_{\beta\to\infty} \frac{1}{\beta} \log \nu_\beta(O) = -\inf_{x\in O}\{I(x)\},$$

and,
(2) for any closed set $F \subset \mathbb{T}^n$,

$$\limsup_{\beta\to\infty} \frac{1}{\beta} \log \nu_\beta(F) = -\inf_{x\in F}\{I(x)\}.$$

It follows from Varadhan's Integral Lemma (Sect. 4.3 in [8]) that, for any C^∞ function $F(x)$,

$$\lim_{\beta\to\infty} \frac{1}{\beta} \log \int e^{\beta F(x)}\, d\nu_\beta(x) = \sup_{x\in\mathbb{T}^n}\{F(x) - I(x)\}.$$

The W^t_β-Kernel is defined by

$$e^{W^t_\beta(y,x)} = \int e^{-\beta\int_0^t V(\alpha(s))\,ds - \beta\int_y^x w}\, \mathscr{W}^{\beta^{-1}}_{y,x,t}(d\alpha) =$$

$$\int e^{-\beta\int_0^t V(\alpha(s))\,ds - \beta < P,(x-y) >}\, \mathscr{W}^{\beta^{-1}}_{y,x,t}(d\alpha).$$

Note the plus sign on V.
Note that we exchange x and y above (with respect to the previous considerations).
It is known (see [1]) that for any β and any t

$$\psi_\beta(x) = \int e^{W^t_\beta(y,x)}\, \psi^*_\beta(y)\, dy = \int e^{W^t_\beta(y,x)} \frac{1}{\psi_\beta(y)} d\nu_\beta(y)$$

Now from Schilder's Theorem and Varadhan's Integral Lemma (see [8] also Theorem 4.3.9 in [2])

$$-W(y,x) := -\lim_{\beta\to\infty} \frac{1}{\beta} \log e^{W^{\frac{1}{\beta}}_\beta(y,x)} =$$

$$\inf_{\alpha\in C^1([0,1],\mathbb{T}^n),\,\alpha(0)=y,\,\alpha(1)=x} \left\{ \int_0^1 [-V(\alpha(s) + w_{\alpha(s)}(\alpha'(s))]\, ds + \int_0^1 \frac{1}{2}\, ||\alpha'(t)||^2 \right\}$$

Note above the plus sign on w.
The function $W(y,x)$ is the function $-I(y,x)$ in the notation of [2].
For any β

$$\frac{1}{\beta}\log(\psi_\beta(x)) = \frac{1}{\beta}\log\left(\int e^{W^{\frac{1}{\beta}}_\beta(y,x)}\, \psi_\beta(y)^{-1}\, d\nu_\beta(y)\right).$$

Taking limits as $\beta \to \infty$ and using the Varadhan's integral Lemma once more we get

$$-u(x) = \sup_{z \in \mathbb{T}^n} \{W(z, x) + u(z) - I(z)\}.$$

Therefore, for any x, y we get that

$$-u(y) - u(x) \ \geq \ W(y, x) - I(y).$$

From this we get

Proposition 2

$$u(y) + u(x) \ \leq \ -W(y, x) + I(y) \ = \ c(y, x).$$

and the pair (u, u) is (-W+I)-admissible.

In the same way the pair (u, u^*) is -W-admissible.

Proposition 3 *If $\hat{\eta}$ is an optimal minimizing transport probability for c and if $(f, f^{\#})$ is an optimal pair in $\mathscr{F}$, then the support of $\hat{\eta}$ is contained in the set*

$$\{\, (x, y) \in M \times M \ \textit{such that} \ (f(x) - f^{\#}(y)) \ = \ c(x, y) \,\}.$$

Proof It follows from the primal and dual linear programming problem formulation. The condition above is called the complementary slackness condition (see [9]). □

If one finds $\hat{\eta}$ an an admissible pair $(f, f^{\#})$ satisfying the above claim (for the support) one solves the Kantorovich problem, that is, one finds the optimal transport probability $\hat{\eta}$.

From the above it follows.

Proposition 4 *For (x, y) in the support of $\hat{\mu}$ we have*

$$u(x) + u(y) \ = \ -W(y, x) + h(y, y) = c(x, y),$$

or

$$u(x) + u(y) \ = \ -W(y, x) \ + \ (u^*(y) + u(y)).$$

This means, for (x, y) in the support of $\hat{\mu}$

$$u(x) - u^*(y) \ = \ -W(y, x)\,.$$

In other words, for any x, y in the support of $\hat{\mu}_{min}$ we have that $u(x)$ is given by

$$\inf_{\alpha \, \in C^1([0,1], \mathbb{T}^n), \, \alpha(0) = y, \, \alpha(1) = x} \left\{ \int_0^1 \, [\, - V(\alpha) + w(\alpha')\,]\, ds \ + \ \int_0^1 \, \frac{1}{2} \, ||\alpha'||^2 ds \right\} \ + \ u^*(y).$$

References

1. Anantharaman, A.: Entropie et localisation des fonctions propres, Document de Synthese, Ecole Normal Superieure de Lyon, France, (2006). http://www.math.polytechnique.fr/~nalini/HDR.pdf
2. Anantharaman, N.: Counting geodesics which are optimal in homology. Ergod. Theory Dyn. Syst. **23**(2), 353–388 (2003)
3. Anantharaman, N.: On the zero-temperature or vanishing viscosity limit for Markov processes arising from Lagrangian Dynamics. J. Eur. Math. Soc. **6**(2), 207–276 (2004)
4. Baraviera, A., Lopes, A.O., Thieullen, Ph: A large deviation principle for equilibrium states of holder potentials: the zero temperature case. Stoch. Dyn. **6**, 77–96 (2006)
5. G. Contreras, R. Iturriaga, *Global minimizers of autonomous La-grangians*, 22° Colóquio Brasileiro de Matemática, IMPA, 1999
6. Contreras, G., Lopes, A.O., Oliveira, E.R.: Ergodic transport theory, periodic maximizing probabilities and the twist condition. In: Zilberman, D., Pinto, A. (eds.) Modeling, Optimization, Dynamics and Bioeconomy I. Springer proceedings in mathematics and statistics, pp. 183–219 (2014)
7. Davies, E.: Spectral Theory and Differential operators. Cambridge Press, Cambridge (1995)
8. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Springer, Berlin (1998)
9. Evans, L., Gomes, D.: Linear programming interpretation of Mather's variational principle. ESAIM Control Optim. Calc. Var. **8**, 693–702 (2002)
10. Fathi, A.: The Weak Kam Theorem in Lagrangian Dynamics. Cambridge studies in advanced mathematics. Cambridge University Press, UK (2008)
11. Fathi, A.: Théorème KAM faible et théorie de Mather sur les systèmes lagrangiens. Comptes Rendus de l'Académie des Sciences, Série I, Mathématique **324**, 1043–1046 (1997)
12. Gomes, D., Lopes, A.O., Mohr, J.: Wigner measures and the semi-classical limit to the Aubry-Mather measure. Appl. Math. Res. Express **2012**(2), 152–183 (2012)
13. Levitan, B.M., Sargsjan, I.: Sturm-Liouville and Dirac Operators. Kluwer, Dordrecht (1991)
14. Lopes, A.O., Oliveira, E.O., Thieullen, Ph: The dual potential, the convolution kernel and transport in ergodic optimization. In: Bourguignon, J.-P., Jelstch, R., Pinto, A., Viana, M. (eds.) Dynamics, Games and Science. International Conference and Advanced School Planet Earth DGS II, Portugal (2013), pp. 357–398. Springer, Berlin (2015)
15. Lopes, A.O., Mengue, J.: Duality theorems in Ergodic transport. J. Stat. Phys. **149**(5), 921–942 (2012)
16. Lopes, A.O., Oliveira, E.R., Smania, D.: Ergodic transport theory and piecewise analytic subactions for analytic dynamics. Bull. Braz. Math. Soc. **43**(3), 467–512 (2012)
17. Lopes, A.O., Mengue, J.K., Mohr, J., Souza, R.R.: Entropy and variational principle for one-dimensional lattice systems with a general a-priori measure: positive and zero temperature. Ergod. Theory Dyn. Syst. **35**(6), 19251961 (2015)
18. Lopes, A.O., Oliveira, E.R.: On the thin boundary of the fat attractor (2012)
19. Lopes, A.O., Ruggiero, R.: Large deviations and Aubry-Mather measures supported in nonhyperbolic closed geodesics. Discret. Contin. Dyn. Syst. A. **29**(3), 1155–1174 (2011)
20. Rachev, S., Ruschendorf, L.: Mass Transportation Problems, vol. I. Springer, Berlin (1998)
21. Villani, C.: Topics in Optimal Transportation. AMS, Providence (2003)
22. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2009)

Dynamics of a Fixed Bed Adsorption Column in the Kinetic Separation of Hexane Isomers in MOF ZIF-8

Patrícia A. P. Mendes, Alírio E. Rodrigues, João P. Almeida and José A. C. Silva

Abstract A fixed bed adsorption mathematical model has been developed to describe the kinetic separation of hexane isomers when they flow through a packed bed containing the microporous Metal-Organic Framework (MOF) ZIF-8 adsorbent. The flow of inert and adsorbable species through the fixed bed is modeled with fundamental differential equations according to the mass and heat conservation laws, a general isotherm to describe adsorption equilibrium and a lumped kinetic mass transfer mechanism between bulk gas phase and the porous solid. It is shown that a proper combination of two characteristic times (the residence time of the gas in the fixed bed, τ_{fb} and the characteristic time of diffusion of solutes into the pores τ_{dif}) can lead to very different dynamics of fixed bed adsorbers where in a limiting case can gives rise to a spontaneous breakthrough curves of solutes. The numerical simulations of an experimental breakthrough curve with the developed mathematical model clearly explain the complete separation between linear n-Hexane (nHEX) and the respective branched isomers: 3-Methyl-Pentane (3MP) and 2, 2-Dimethyl-Butane (22DMB). The separation is due to significant differences in the diffusivity parameters τ_{dif} between 3MP and 22DMB and the residence time of the gas mixture τ_{fb} within the fixed bed. This work shows the importance of mathematical modelling for the comprehension and design of adsorption separation processes.

P. A. P. Mendes · J. A. C. Silva (✉)
Portugal and LSRE - Laboratory of Separation and Reaction Engineering,
Departamento de Engenharia Química, Escola Superior de Tecnologia e Gestão,
Instituto Politécnico de Bragança, S/N, 5301-857 Bragança, Portugal
e-mail: jsilva@ipb.pt

A. E. Rodrigues
LSRE - Laboratory of Separation and Reaction Engineering,
Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias, S/N, Porto, Portugal

J. P. Almeida
CeDRI-IPB and LIAAD - INESC TEC and Escola Superior de Tecnologia e Gestão,
Instituto Politécnico de Bragança, 5301-857 Bragança, Portugal
e-mail: jpa@ipb.pt

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_12

1 Introduction

The gasoline used by automotive cars has a high Research-Octane Number (RON above 95) and is mainly obtained from a stream fraction of the crude distillation column called Light Naphtha which in turn has a very low RON content (average RON of 63). To be upgraded the Light Naphtha mainly composed by n-pentane (RON 61.7) and n-hexane (RON 24.8) is treated in processes such as: Hysomer, TIP or IPsorb [1-5] where a catalytic reactor isomerize the linear compounds into branched ones with a much higher RON content. The final output stream of TIP processes (RON around 88) needs to be further upgraded with additives. Some of these additives were forbidden due to their toxicity (the case of lead) to make it usable by cars. Also they increase significantly the final price of gasoline (MTBE and ETBE). Since this final TIP output stream still contains compounds such as the linear nHEX (RON 24.8), mono-branched 3MP (RON 74.5) and di-branched 22DMB (RON 94.0), there is nowadays a significant effort to discover new adsorbents to separate in a clean way the low from the high RON paraffins of the final TIP output stream (the separation by distillation is very energy consuming due to the close boiling point of the isomers) to reduce the use of additives.

Among the several adsorbents being discovered for the separation of close boiling point compounds, the zinc methyl-imidazolate ZIF-8 [6] with the sodalite (SOD) topology, which possesses a significant porosity ($S_{BET} \sim 1800\ m^2g^{-1}$; $V_p \sim 0.66\ cm^3g^{-1}$) involving large spherical cavities (11.4 Å) connected by a flexible six members rings of free aperture of 3.4 Å, is of particular interest due to its chemical and thermal robustness. Regarding the separation of pentane/hexane isomers in ZIF-8, the study of Luebbers et al. [13] showed a complete separation between the linear n-pentane and the branched isopentane. More recently, a screening study of Peralta et al. [17] proved that ZIF-8 is an interesting material for the separation of linear nHEXand branched hexane isomers. However, under static conditions, Ferreira et al. [7] showed that the linear and mono-branched hexane isomers were adsorbed well, but 22DMB was totally excluded. A similar study was performed by Zhang et al. [23] obtaining completely different results. All hexane isomers were adsorbed in ZIF-8 especially the branched ones. However, they estimated diffusional parameters from the uptakes and found that the diffusion selectivities for nHEX/3MP, nHEX/23DMB and 3MP/23DMB were of 20, 54, and 3, respectively. They concluded that the separation linear/mono/di-branched hexane isomers could also be attained by a kinetic selectivity.

Mathematical modelling is a valuable tool in the design and optimization of industrial processes. Most of the industrial adsorption processes occur in fixed beds and it is the overall dynamics of this packed bed system, rather than the adsorption equilibrium or the adsorption kinetics in a single particle, that determine the efficiency of such process [11]. Fundamentals of mathematical modelling of fixed bed adsorbers are presented by Ruthven [19] and Yang [22].

The aim of this work is the formulation of a mathematical model to simulate the transient adsorption behavior of a mixture of hexane isomers (nHEX, 3MP, 22DM-

B) flowing through a packed bed containing the microporous adsorbent ZIF-8. The physical parameters of the model are determined on the basis of results of equilibrium and kinetics of adsorption and from correlations available in the literature. The numerical solution of the coupled mass and heat balances partial differential equations is obtained by orthogonal collocation. Application of the model is illustrated by a typical example, a so called breakthrough curve, which shows how the different hexane isomers are separated from one another.

2 Simulation of Fixed Bed Adsorption Dynamics

2.1 Mathematical Model

Consider a fixed bed adsorption column of length L, void fraction ε_b, packed with an adsorbent through which a fluid mixture of hexane isomers flows at a molar flow rate F. Let C represent the total gas concentration of all species in the fluid mixture and $\bar{q}_i$ the average adsorbed concentration of absorbable species i in the solid phase. The total material balance in a section between axial planes z and $z + \Delta Z$ from the entrance of the bed over a period of time t to $t + \Delta t$ yields, in the limit, the following first order partial differential equation,

$$\frac{\partial F}{\partial z} + \varepsilon_b \frac{\partial C}{\partial t} + (1 - \varepsilon_b) \sum_{i=1}^{ncp} \frac{\partial \bar{q}_i}{\partial t} = 0. \tag{1}$$

The initial and boundary conditions for a clean column subjected to a step change of adsorbable species at the inlet and at time zero are,

Boundary condition

$$z = 0; \quad t > 0; \quad F = F_f. \tag{2}$$

Initial condition

$$t = 0, \forall z; \quad \bar{q}_i = 0; \quad F = F_f; \quad C = C_f, \tag{3}$$

where the subscript f represents the feed conditions.

The differential fluid phase mass balance for a solute species i represented by an axially dispersed plug flow pattern is the second order partial differential equation,

$$-\varepsilon_b D_{ax} \frac{\partial}{\partial z}\left(C \frac{\partial y_i}{\partial z}\right) + \frac{\partial (F y_i)}{\partial z} + \varepsilon_b \frac{\partial (C y_i)}{\partial t} + (1 - \varepsilon_b) \frac{\partial \bar{q}_i}{\partial t} = 0. \tag{4}$$

where y_i is the molar gas fraction of solute i and D_{ax} is the coefficient of axial dispersion. The initial and boundary conditions (known as Danckwerts boundary conditions [5]) are,

Initial condition

$$t = 0, \forall z; \quad y_i = \bar{q}_i = 0. \tag{5}$$

Boundary conditions

$$z = 0; \quad t > 0; \quad Fy_{if} = Fy_i - \varepsilon_b D_{ax} C \frac{\partial y_i}{\partial z} \tag{6}$$

$$z = L; \quad t > 0; \quad \frac{\partial y_i}{\partial z} = 0, \tag{7}$$

where L is the length of the column.

Due to axial diffusion, the molar fraction of the adsorbable entering the bed is different from the one in the entrance. Accordingly, Eq. (5) ensures that the mass fed to the column is equal to the one that crosses the plane at $z = 0$. At the outlet of the bed ($z = L$), Eq. (6) simply assumes that the concentration gradient ends and the molar fraction of the absorbable just at the end of the bed is not affected by counter-diffusion.

The mass transfer rate from bulk fluid phase to solid particles is mainly governed by: (i) external fluid film resistance around the particles, and (ii) intraparticle diffusion of solutes. A rigorous treatment of intraparticle diffusion of solutes leads to a diffusion model with partial differential equations that incorporates several mechanisms and a new radial coordinate (if adsorbents particles are spherical or cylindrical). To simplify the solution, a linear rate model is generally used,

$$\frac{\partial \bar{q}_i}{\partial t} = k_{LDF}(q^* - \bar{q}_i), \tag{8}$$

where q^* is the adsorbed phase concentration of species i in equilibrium with gas phase concentration, $\bar{q}_i$ is the average adsorbed phase concentration of species i within the particle and k_{LDF} is called the Linear Driving Force (LDF) mass transfer coefficient. Glueckauf [9] showed that the parameter k_{LDF} for spherical particles is equal to $15D_c/r_c^2$ where D_c is the diffusivity constant and r_c the particle radius.

Adsorption is an exothermic *phenomenum* and the importance of heat effects should also be considered in the modelling of an adsorption column. Consider a non-isothermal, non-adiabatic column with axial heat dispersion. Let T be the temperature in bulk gas phase, T_s the temperature of solid phase, T_w the temperature of the surroundings, c_{pg} the heat capacity per unit mol of gas, K_{ax} the axial heat dispersion coefficient, h_p the heat transfer coefficient between gas and solid phase, h_w the overall heat transfer coefficient at the wall of the column, a_c the specific area of the column and a_p the specific area of the particle. Then, the following differential energy balance may be formulated to give the equation

$$-K_{ax}\frac{\partial^2 T}{\partial z^2}+Fc_{pg}\frac{\partial T}{\partial z}+\varepsilon Cc_{pg}\frac{\partial T}{\partial t}+(1-\varepsilon_b)a_p h_p(T-T_s)+a_c h_w(T-T_w)=0. \quad (9)$$

Boundary conditions

$$z=0;\quad t>0;\quad Fc_{pg}T_f=Fc_{pg}T-K_{ax}\frac{\partial T}{\partial z} \quad (10)$$

$$z=L;\quad t>0;\quad \frac{\partial T}{\partial z}=0. \quad (11)$$

Initial condition

$$t=0,\forall z;\quad T=T_s=T_f. \quad (12)$$

Since mass and heat transfer are similar mechanisms, the previous boundary conditions are applied by analogy with the Danckwerts boundary conditions.

The existence of an interphase heat transfer mechanism within the column implies that in certain cases the temperature of bulk fluid phase T is different from the solid phase temperature T_s, under transient conditions. The energy balance for the solid phase neglecting radial temperature gradients is

$$c_{ps}\frac{\partial T_s}{\partial t}=a_p h_p(T-T_s)+\sum_{i=1}^{ncp}(-\Delta H_i)\frac{\partial \bar{q}_i}{\partial t}, \quad (13)$$

where, c_{ps} is the heat capacity per unit volume of solid and $(-\Delta H_i)$ the heat of adsorption per mole of solute species i or, in other words, the amount of heat that is generated by adsorption within the particle.

2.2 Adsorption Isotherm

The equilibrium thermodynamic relation between adsorbed phase concentration of solutes q_i and the respective gas phase concentration c_i can be represented by the ideal localized model introduced by Langmuir [12]

$$q_i=\frac{q_m K_i c_i}{1+\sum_{j=1}^{ncp}K_j c_j}, \quad (14)$$

where q_m is supposed to be constant and independent of temperature, in order to give thermodynamic consistency to the model, and K_i a Langmuir isotherm constant. The Langmuir isotherm constant K_i is strongly dependent upon temperature obeying the Van't Hoff dependence

$$K_i = K_i^0 e^{-\Delta H_i/RT}, \tag{15}$$

where K_i^0 is the frequency factor of the Langmuir constant, R is the universal gas constant and T is the temperature.

2.3 Numerical Solution of Model Equations

The fluid flow problem formulated above consist in a set of partial differential equations (PDES) subjected to boundary and initial value conditions that lead to parabolic equations which are time dependent. Several numerical techniques have been developed to solve PDEs where the most widely used are the: (i) finite-difference (FD); (ii) the method of lines (MOL) [20] and (iii) orthogonal collocation (OC) [8]. The choice between the methods depends on the complexity, simplicity in setting-up the numerical solution, accuracy, stability and computational effort [2, 18].

The classical finite difference method consists in replacing the derivatives in the PDE by finite difference approximations using a uniform mesh. The method of lines is similar to the finite-difference except we do not discretize the time variable. In the orthogonal collocation technique trial functions are chosen as sets of orthogonal polynomials and collocation points are the roots of these polynomials.

The numerical methods convert the PDES into a set of non-linear equations (FD) or in set of ordinary's differential equations (ODES) that must be solved by a proper numerical technique. The diffusion-adsorption problem formulated above can give rise to profiles very steep with the necessity to increase the mesh if we use the FD or MOL methods or the roots of the polynomials in the OC method, to obtain stability and accuracy. Consequently the numerical computation effort increase significantly to obtain accuracy in such cases. The most widely used methods to handle the diffusion-adsorption problems are the MOL and OC techniques. The flexibility of the OC technique in handle such problems, for example using orthogonal polynomials [21], means that the solution error decreases faster as the polynomial order increases. However, the selection of a method to handle the solution of a PDE problem is in most cases a matter of experimentation since many factor have to be considered [2].

Based on such experimentation (during past years) and given the collection of subroutines by Villdasen and Michelsen [21] to help readers handle the solution of PDE problems by using the accurate computation of collocation points obtained by the zeros of an orthogonal polynomial, $P_{N^{(\alpha,\beta)}}(x)$, called Jacobi polynomial, are the base to select the OC technique to obtain a solution for the problem formulated above.

In this work, the set of coupled partial differential equations was reduced firstly to a set of ordinary differential/algebraic equations (DAE's) applying orthogonal collocation technique to the spatial coordinate [16]. In this reduction, the first and second order differential terms were replaced by collocation matrices $A(i, j)$ and $B(i, j)$, respectively. The collocation points were given by the zeros of Jacobi polynomials $P_{N^{(\alpha,\beta)}}(x)$, with $\alpha = \beta = 0$, calculated by subroutine JCOBI. The collocation

matrices $A(i, j)$ and $B(i, j)$ were found by using subroutine DFOPR. A FORTRAN code of both subroutines can be found in Villadsen and Michelsen's book [21]. The number of interior interpolation points N was chosen to give stability to the numerical solution of discretized equations. The resulting system was solved using a fifth order Runge-Kutta code (ODE's) together with Gauss elimination for the algebraic equations. Sixteen collocation points appeared to give satisfactory accuracy for all calculations performed. For two adsorbable species, this results in 128 (64 × 2) ODEs being integrated at the same time: 32 (16 × 2) from the Mass balance to adsorbable species; 32 (16 × 2) from the equation representing the Mass transfer rate, 32 (16 × 2) from the energy balance in the gas phase and 32 (16 × 2) from the energy balance for the solid phase. At the same time there are 32 (16 × 2) equations being solved by Gaussian elimination from the equation representing the overall mass balance.

3 Experimental Framework

3.1 Adsorbent and Sorbates

ZIF-8 crystals were synthesized for the adsorption studies of hexane isomers. For that, a solution of 5 g of 2-methylimidazole (H-MeIM; 60.9 mmol; Alfa Aesar, 97%) in 25 mL of methanol (6.15 mmol; VWR, 99.9%) was poured into a solution of 2.305 g zinc nitrate hexahydrate, ($Zn(NO_3)_2.6H_2O$; 7.5 mmol; Aldrich, 99%) in 25 mL methanol (6.15 mmol, VWR, 99.9%). The mixture was then putted into a metallic PAAR digestion bomb at 100 C during 16 h. The resulting white powder was filtered and washed with ethanol. This procedure was repeated five times to finally obtain the enough ZIF-8 amount for the fixed bed experiments (∼2 g). The crystallinity of each batch was checked by XRPD before mixing all of them. Thereafter the powder was characterized by X-ray powder diffraction patterns were collected in a SIEMENS D5000 diffractometer (θ–2θ), Thermogravimetric analysis using a Perkin Elmer Diamond TGA/DTA STA 6000. The BET surface area was 1950 $m^2.g^{-1}$ with a total pore volume around 0.66 $cm^3.g^{-1}$. The hexane isomers used in the experimental study were obtained from Merck (Germany) and are all above 99% purity.

3.2 Single and Multicomponent Fixed Bed Experiments

The experimental data were obtained in an apparatus developed for the measurement of single and multicomponent breakthrough curves consisting of three main sections. The preparation section includes a syringe pump used to introduce the adsorbable species in the carrier gas followed by a heating chamber where this stream is com-

pletely vaporized. The adsorption section consists in a 4.6 mm *i.d.* stainless steel column with 80 mm in length containing the adsorbent and placed in a ventilated chromatographic oven, as well as a heated collector to collect samples at the outlet of the column. The third part is an analytical section composed by a chromatographic column and a flame ionization detector (FID). Complete information about the experimental setup is reported in, for example, [1].

4 Results and Discussion

The mathematical model developed contains important time dependent group parameters that influence the overall dynamics of the fixed bed namely: (i) the residence time (contact time or space time) measured by a characteristic time L/v_i; and (ii) the Linear Driving Force (LDF) mass transfer coefficient k_{LDF} equal to $15D_c/r_c^2$. Here, L/v_i has units of time and the group $15D_c/r_c^2$ has units of the reciprocal of time. The term L/v_i can be viewed as the residence time τ_{fb} of the gas in the column, that is, the time that gas and the adsorbent within the fixed bed are in contact and the reciprocal of $15D_c/r_c^2$ as the time τ_{dif} that the adsorbable species spent needs to diffuse into the pores of the adsorbent. The importance of these parameters in the overall dynamics of the fixed bed will be analyzed.

Figure 1 shows the effect of changing τ_{dif} for a constant value of $\tau_{fb} = 1$ min under isothermal conditions. The parameters of the model used in simulations are shown in Table 1. When τ_{dif} is 10 times lower ($\tau_{dif} = 0.1$ min) the diffusion limitations are insignificant when compared to the residence time of the gas in the bed and the breakthrough curve is a pure "shock wave". When τ_{dif} is ten times higher (that is, the time that solutes need to diffuse into the pores is much higher than the residence time in the bed) the breakthrough curves become dispersive (e.g. $\tau_{dif} = 10$ min) indicating strong mass transfer resistance. In the worst case presented ($\tau_{dif} = 100$ min), the solute breakthroughs spontaneously at the residence time of the bed because the time it spends in the bed is shorter than the time it needs to diffuse into the pores. These results mean that a proper combination of τ_{dif} and τ_{fb} can lead to very different dynamics and consequently breakthrough curves of fixed bed adsorbers.

Previous studies of hexane isomers adsorption in the microporous ZIF-8 are controversial. Peralta et al. [17] show in a breakthrough apparatus that 3MP was kinetically separated from 22DMB. In contrast, Chang et al. [3] and Luebbers et al. [13] proved that branched alkanes can be totally sieved from the linear ones. Moreover, results of Ferreira et al. [7], who studied the same system but in static conditions (manometric system coupled with a micro calorimeter), are different from both of the previously mentioned ones. They noticed that the linear and mono-branched isomers were well adsorbed, but the di-branched 22DMB was totally excluded. Moreover, the recent work of Zhang et al. [23] that studied the same system in static conditions (uptake system), shows completely different results relatively to the results obtained in the experiments of Ferreira et al. [7]. In this case [23], all isomers can enter into the pores of ZIF-8 and the amounts adsorbed of the branched isomers are even

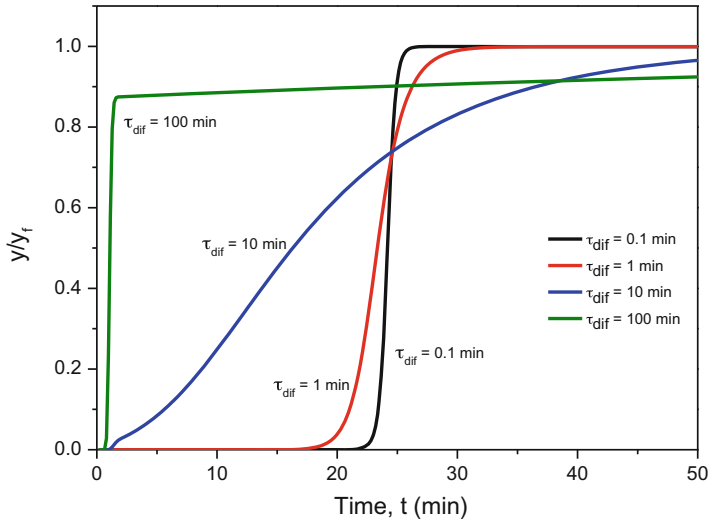


Fig. 1 Effect of changing the characteristic time of diffusion τ_{dif} in the overall dynamics of an isothermal fixed bed adsorber for a constant value of the residence time of the gas τ_{fb} equal to 1 min

Table 1 Model parameters for the simulation of breakthrough curves shown in Fig. 1

Isotherm parameters (nHEX)	
q_m	1 mmol/g
K_i	0.01 kPa^{-1} at 313 K
Experimental conditions	
C_f	40 mol/m^3
Q	20 mL/min
P_c	101.3 kPa
V_c	40 cm^3
y_{if}	0.5
Model parameters	
Pe	100
Greek letters	
ε_b	0.5
ρ_b	0.7 g/cm^3

higher than the linear ones. However, they estimated diffusional parameters from the uptakes and found that the diffusion selectivities for nHEX/3MP, nHEX/23DMB and 3MP/23DMB were of 20, 54, and 3 respectively in ZIF-8. This means that the separation linear/ branched can be kinetically driven.

A convenient way to evaluate the performance of an adsorbent for a certain separation is to perform a set of experimental breakthrough curves in a packed column. These response curves to a step change in solute concentration at inlet contain the

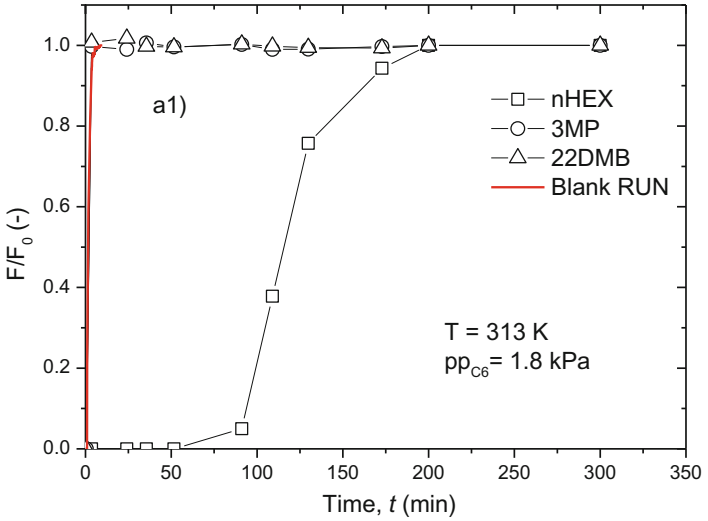


Fig. 2 Typical experimental breakthrough curve for sorption of an equimolar ternary mixture nHEX/3MP/22DMB in ZIF-8. Experimental conditions: a1) pp = 1.8 kPa, T = 313 K; Red lines represent blank runs at the same experimental conditions with the column filled with glass spheres

basic information of the underlying phenomena that governs the system. With such data is it possible using a convenient mathematical model to identify which are the controlling mechanisms of the process and thereafter to proceed to a convenient design of an industrial process. Moreover, it is generally the dynamics of the fixed bed system, rather than the equilibrium conditions, and then kinetic mechanism in single particles that control the adsorption separation process.

To clarify the discrepancies previously observed by several authors, we have performed several breakthrough experiments with the mixture nHEX/3MP/22DMB in commercial and home-made ZIF-8 [14]. Figure 2 shows a typical breakthrough curve where it can be seen that branched isomers 22DMB and 3MP practically elute the column at the beginning of the experiment with no separation between them, but they are completely separated from linear nHEX that elute much later. This result indicates that 3MP and 22DMB leave the column practically with no adsorption in ZIF-8 in contrast with nHEX that adsorbs significantly in the column. These results are controversial since Zhang et al. [23] observed in a static system that all hexane isomers adsorb on ZIF-8. However, they are in agreement with the ones found by Chang et al. [3] and Luebers et al. [13] who studied the same system in a chromatographic column. A possible explanation could be the different methods used to study the adsorption of hexane isomers in ZIF-8. Some studies were performed in flow conditions and others in static conditions. However, that justification is not totally satisfactory since Peralta et al. [17] performed also the study in a breakthrough apparatus and found a certain degree of separation between 3MP and 22DMB. A possibility concerning this discrepancy could be related to differences in residence time of the gas mixture in the fixed bed column. In Peralta et al. experiments, the

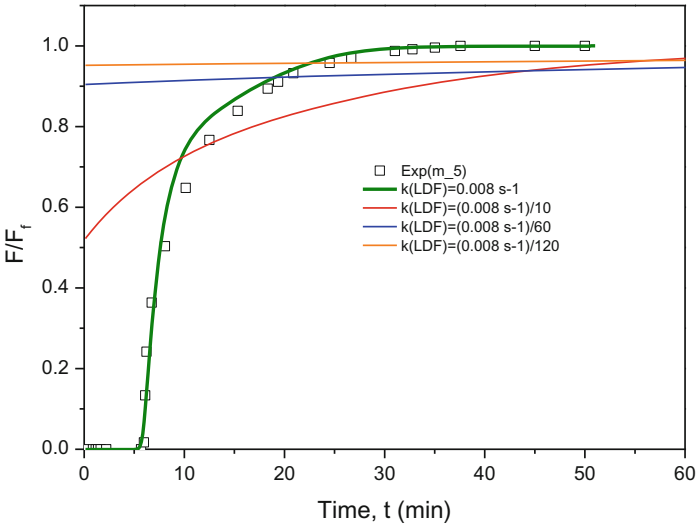


Fig. 3 Single component nHEX breakthrough at 313 K. The green line represents the simulation of the experiment with a K_{LDF} of 0.008 s^{-1}. The other lines (red, blue and orange) show the effect of decreasing the K_{LDF} by 10, 50 and 100 times, respectively, in the breakthrough curve (or increasing τ_{dif}). Points are experimental data and lines are model predictions

contact time (residence time) measured by $\tau_{fb} = L/v_i$ is around 20 s, where v_i is the interstitial velocity. In our experiments, it ranges from 1 to 4 s. Since 3MP and 22DMB are slow diffusing species [23], this could explain why 3MP enters partially in the adsorbent in the case of Peralta's experiments due to at least 5 times higher residence time of the gas mixture if the fixed bed column. Apart from being slow diffusion species, the ratio of diffusivities 3MP/23DMB is 3 which combined with the ratio of diffusivities and residence time of the gas in the column could support the separation observed in the breakthrough curves.

To prove that the separation between nHEX and the branched isomers 3MP and 22DMB observed in our experiments is kinetically driven, we use the mathematical model previously described to simulate a single component breakthrough. For the study, we select an experimental single component breakthrough curve of n-HEX since we do not have isotherm data for 3MP and 22DMB because they do not adsorb in the packed column. Figure 3 shows the experiment together with the simulations performed by the mathematical model. The experimental conditions and model parameters used for the simulation are shown in Table 2. For the simulation we use the corrected diffusivity D_0 for nHEX reported by Zhang et al. [23] which is 3.4×10^{-19} m^2/s at 313 K. Since this value is a corrected diffusivity we transform it to the Fickian diffusivity using the correction factor for the Langmuir isotherm $D_c = D_0/(1 - q/q_{max})$ (see [19]). Since the amount adsorbed in the experiment is 2.94 mmol/g and the total loading taken from the Langmuir isotherm is 3.13 mmol/g (see [14]) we estimate a Fickian diffusivity D_c of 5.4×10^{-18} m^2/s for nHEX. If we use the LDF model for the mass transfer in the bed, the characteristic

Table 2 Experimental conditions and model parameters for the simulation of the single component (nHEX) breakthrough experiment shown in Fig. 3

Isotherm parameters (nHEX)	
q_m	3.13 mmol/g
K_i	1.73 kPa^{-1} at 313 K
Experimental conditions	
C_f	57.6 mol/m^3
F_f	1.33 $mol/m^2.s$
m_a	0.318 g
P_c	150 kPa
T_f	313 K
y_{if}	0.0656
Model parameters	
c_{pg}	20 J/mol.K
c_{ps}	1.6 J/g.K
d_p	0.2×10^{-6} m
d_c	4.6×10^{-3} m
D_{ax}	2.59×10^{-5} m^2/s (see *Note 1*)
h_p	42500 $W/m^2.K$ (see *Note 2*)
h_w	0.38 $W/m^2.K$ (see *Note 3*)
k_{LDF}	0.008 s^2 (see *Note 4*)
K_{ax}	0.030 W/m.K
L	0.08 m
Greek letters	
ε_b	0.5

Note 1 Calculated by the correlation $D_{ax} = 0.7D_m + 0.5d_p v$, taken from Ruthven [19]. The axial mass Peclet number is 179

Note 2 This value was estimated from the limit of $Nu = 2$ and it can be considered a very high value, which means that the temperature between solid and bulk gas phase is in equilibrium

Note 3 This parameter was obtained through the fitting of the experimental breakthrough curve

Note 4 The LDF parameter was calculated by $k_{LDF} = 150D_c/r_c^2$, with D_c calculated from the data of Zhang et al. [23]

mass transfer coefficient will be $K_{LDF} = 15 \times 5.4 \times 10^{-18}/(0.1 \times 10^{-6})^2 = 0.008$ s^{-1} (we assume that the size of the crystals is similar to size of the commercial ones, that is 0.2 μm [17]). The mean residence time or contact time $\tau_{fb} = L/v_i$ for the experiment is around 1.3 s. Figure 3 shows the simulation of the breakthrough of nHEX (green line) using the mass transfer parameter $k_{LDF} = 0.008$ s^{-1}. The experimental conditions and model parameters are specified in Table 2. To fit the curvature of the breakthrough curve we also need to use a non-isothermal model due to the slow approach of concentration to equilibrium. It is clear from Fig. 3 that it is possible to fit the profile of the nHEX curve with the diffusivity data reported by Zhang et al. [23] and using an LDF model. To see what happens in the fixed bed by decreas-

ing the diffusivity (or increasing τ_{dif}) we simulate the same experiment using a value of k_{LDF} 10, 60 and 120 times lower (we note that Zhang et al. [23] report a diffusivity 20 times lower for 3MP and 147 times lower for 23 DMB at 313 K). Figure 3 clearly shows that if we decrease at least 10 times the diffusivity of the adsorbable species in the bed the compound will come out of the column immediately at the residence time τ_{fb}. This simulation clearly explains that the separation between normal and branched paraffins is indeed kinetically driven and can be predicted with diffusivity data already published using a convenient mathematical model.

These results also prove that the discrepancy between experiments in flow and static systems are due to kinetic considerations proper of fixed bed adsorbers. In a static system, the gas mixture of hexane isomers contact with the adsorbent ZIF-8 in a completely different time scale compared to experiments that could occur for few seconds. This means that an appropriate selection of the residence time of the gas mixture τ_{fb} in the fixed bed could give rise to a separation of hexane isomers due to their different diffusivities apart for being all adsorbed with identical amounts.

Figure 3 also shows that to capture the profile of the breakthrough curve it is necessary to use a non-isothermal model. Figure 4 shows the simulation results for the concentration and temperature profiles of the breakthrough curves in isothermal and adiabatic cases. The experimental conditions and model parameters are specified in Table 2. It is clear that only using a non-isothermal model it is possible to capture the profile of the experimental breakthrough curve. Moreover, the gas temperature

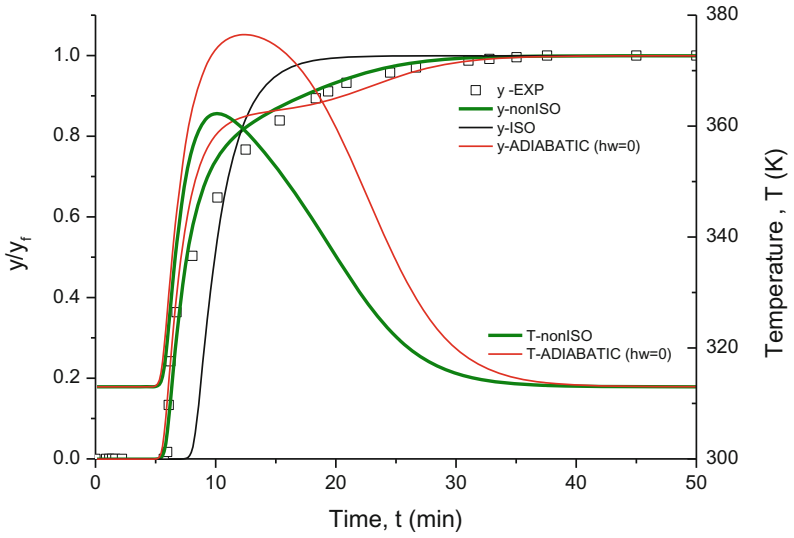


Fig. 4 Simulation of the breakthrough curve shown in Fig. 3 with a non-isothermal, isothermal and adiabatic models. The green lines represent the simulation with the non-isothermal model, the red the adiabatic model, and the black the isothermal model. Points are experimental data and lines are model predictions

of nHEX can increase 50 C at the outlet of the bed due to the large amount of adsorbed nHEX. Figure 4 also shows, for comparison, the profile of composition and temperature in the bed for the limiting isothermal and adiabatic cases. In the adiabatic case the temperature can increase around 70 °C.

5 Conclusions

A mathematical model has been developed to study the hexane isomers adsorption in a fixed bed packed with the microporous adsorbent ZIF-8. Simulations of isothermal breakthroughs show that a proper combination of residence time in the bed measured by τ_{fb} and characteristic time of diffusion of the solutes into the pores of the adsorbent measured by τ_{dif} can lead to very different dynamics of fixed bed adsorbers. It was shown that a τ_{dif} 100 times higher than τ_{fb} leads to spontaneous breakthrough curves of solutes in the fixed bed.

Ternary breakthrough experiments with mixtures of nHEX/3MP/22DMB flowing through the packed bed show a complete separation of linear nHEX from the branched paraffins 3MP/22DMB which spontaneously come out of the bed at the beginning of the experiment. These results contradict previously published works [7, 17, 23] but are similar to other results obtained in chromatographic columns [3, 13].

To explain the uncommon experimental sorption behavior observed, the mathematical model developed in this work was used to simulate an experimental breakthrough curve. The parameters of the mathematical model were obtained from different sources. In particular, the diffusivity values of the hexane isomers were taken from the work of Zhang et al. [13].

The simulations results show that the spontaneous breakthrough curves of the branched hexane isomers 3MP and 22DMB are due to the very low diffusivity of these compounds in ZIF-8 when compared to the one in the linear nHEX. Moreover, a proper combination of the residence time in the bed $\tau_{fb} = L/v_i$ and the characteristic time of diffusion $\tau_{dif} = 1/k_{LDF} = r_c^2/15D_c$ of the hexane isomers in ZIF-8 can give rise to a complete separation between normal and branched paraffins.

It can also be retained from this work that experimental results of hexane isomers sorption in ZIF-8 measured under batch equilibrium conditions or flow systems can be completely different. Indeed, it is noteworthy that it is possible to completely separate linear from branched alkanes while equilibrium isotherms show similar amounts adsorbed for all the compounds.

Finally, this work shows the importance of mathematical modelling in the analysis and interpretation of experimental data and the design of adsorption processes.

References

1. Bárcia, P.S., Silva, J.A.C., Rodrigues, A.E.: Ind. Eng. Chem. Res. **45**, 4316–28 (2006)
2. Cameron, I.T., Hangos, K.: Process Modelling and Model Analysis, 1st edn. Academic Press, Cambridge (2001). ISBN: 9780121569310
3. Chang, N., Gu, Z.Y., Yan, X.P.: Zeolitic imidazolate framework-8 nanocrystal coated capillary for molecular sieving of branched alkanes from linear alkanes along with high-resolution chromatographic separation of linear alkanes. J. Am. Chem. Soc. **132**(39), 13645–7 (2010)
4. Cusher, N.A.: UOP TIP and once-through zeolitic isomerization processes. In: Meyers, R.A. (ed.) Handbook of Petroleum Refining Processes, 3rd edn. McGraw Hill, New York (2004)
5. Danckwerts, P.V.: Continuous flow systems. Distribution of residence times. Chem. Eng. Sci. **2**, 1–13 (1953)
6. Deschamps, A., Jullian, S.: Adsorption in the oil and gas industry. In: Wauquier, J.P. (ed.) Petroleum Refining: Separation Processes, vol. 2. Technip, Paris (2000)
7. Ferreira, A.F.P., Mittelmeijer-Hazeleger, M.C., Granato, M.A., Martins, V.F.D., Rodrigues, A.E., Rothenberg, G.: Sieving di-branched from mono-branched and linear alkanes using ZIF-8: experimental proof and theoretical explanation. Phys. Chem. Chem. Phys. **15**, 8795–8804 (2013)
8. Finlayson, B.A.: The method of weighted residuals and variational principles. Soc. Ind. Appl. Math. (SIAM) (2014). ISBN-10: 1611973236
9. Glueckauf, E.: Formulae for diffusion into spheres and their application to chromatography. J. Chem. Soc. **51**, 1540–1551 (1955)
10. Holcombe, T.C.: U.S. Patent 4,176,053 (1979)
11. Holcombe, T.C.: U.S. Patent 4,210,771 (1980)
12. Langmuir, I.: The adsorption of gases on plane surfaces of glass, mica and platinum. J. Am. Chem. Soc. **40**, 1361–1403 (1918)
13. Luebbers, M.T., Wu, T., Shen, L., Masel, R.I.: Trends in the adsorption of volatile organic compounds in a large-pore metal-organic framework, IRMOF-1. Langmuir **26**(13), 11319–29 (2010)
14. Mendes, P.A.P., Rodrigues, A.E., Horcajada, P., Serre, C., Silva, J.A.C.: Single and multicomponent adsorption of hexane isomers in the microporous ZIF-8. Micropor. Mesopor. Mater. **194**, 146–156 (2014)
15. Minkkinen, A., Mank, L., Jullian, S.: U.S. Patent 5,233,120 (1993)
16. Park, K.S., Ni, Z., Côte, A.P., Choi, J.Y., Huang, R., Uribe-Romo, F.J., Chae, H.K.: M. O'Keeffe, O. M. Yaghi. PNAS **103**(27), 10186–91 (2006)
17. Peralta, D., Chaplais, G., Masseron, A.S., Barthelet, K., Pirngruber, G.D.: Ind. Eng. Chem. Res. **51**(12), 4692–4702 (2012)
18. Rice, R.G., Do, D.D.: Applied Mathematics and Modeling for Chemical Engineers. Wiley, New York (1995)
19. Ruthven, D.M.: Principles of Adsorption and Adsorption Processes. John Wiley and Sons, New York (1984)
20. Schiesser, W.E.: Computational Mathematics in Engineering and Applied Science: ODEs, DAEs, and PDEs. CRC Press, Boca Raton (1994)
21. Villadsen, J.V., Michelsen, M.L.: Solution of Differential Equation Models by Polynomial Approximation. Prentice-Hall Inc., Englewood Cliffs, New Jersey (1978)
22. Yang, R.T.: Gas Separation By Adsorption Processes. Butterworth, Stoneham (1987)
23. Zhang, K., Lively, R.P., Zhang, C., Chance, R.R., Koros, W.J., Sholl, D.S., Nair, S.: Exploring the framework hydrophobicity and flexibility of ZIF-8: from biofuel recovery to hydrocarbon separations. J. Phys. Chem. Lett. **4**, 3618–3622 (2013)

A Simulation Model for the Physiological Tick Life Cycle

Nabil Nassif, Dania Sheaib and Ghina El Jannoun

Abstract In this paper and following an approach used by two of the authors in (Nassif, N.R., Sheaib, D. (2009) On spectral methods for scalar aged-structured population models.) [5], we present a mathematical model for the tick life cycle based on the McKendrick Partial Differential Equation (PDE). Putting this model using a semi-variational formulation, we derive a Petrov–Galerkin approximation to the solution of the McKendrick PDE, using finite element semi-discretizations that lead to a system of ordinary differential equations in time which computations are carried out using an Euler semi-implicit scheme. The resulting simulations allow us to investigate and understand the dynamics of tick populations. Numerical results are presented illustrating in a realistic way the basic features of the computational model solutions.

Keywords McKendrick-von Foerster model · Deterministic tick life cycle model · Petrov–Galerkin procedure · Finite element method

Background

Tick-borne diseases (theleriosis, re-lapsing fever, TBE (tick-borne encephalitis) are serious health problems affecting humans as well as domestic animals in many parts of the world. These infections are generally transmitted through a bite of an infected tick, and it appears that most of these infections are widely present in some wildlife species; hence, an understanding of tick population dynamics and its interaction with hosts is essential to understand and control such diseases [3]. Several field observations on tick biology show a huge polymorphism in their biology (prolificity,

N. Nassif (✉)
American University of Beirut, Beirut, Lebanon
e-mail: nn12@aub.edu.lb

D. Sheaib
University of Oklahoma, Norman, Oklahoma, USA
e-mail: dania.sheaib@ou.edu

G. E. Jannoun
Université de Bordeaux, Bordeaux, France
e-mail: ghina.el-jannoun@inria.fr

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_13

mortality rates, phenology). This polymorphism is enhanced during the parasitic stages of the tick (during feeding stages) because of the interaction between the tick and the host (immunity of the host, surface of exposure, biology of the host). This degree of interaction is again more complicated when the tick-borne infections are considered and describing this tick behavior can be made possible by monitoring infested animals and presenting the observations as descriptive results. Nevertheless, understanding and predicting the mechanisms leading to a determined phenology is quite impossible as much as the prediction of the impact of different control actions. For this reason, modeling represents a powerful tool offering the opportunity to counter these difficulties. It simultaneously offers a dramatic decrease of the control costs.

The objective of the present work is to develop a tick biology model specific to Hyalomma detritum species based on field data that appeared in [1]. A variety of approaches have been used to model the tick population with various degrees of complexity. Models often describe in a discrete way the various stages of tick development: egg-larvae-nymph-adult, whether the ticks are attached to hosts, and if disease is part of the model, whether the ticks themselves are infected [2, 4, 10, 11]. Generally, models are written as systems of ordinary differential equations with or without delay where the physiological structure is always modeled in a discrete form. However, the development of the ticks between stages takes time and this time delay cannot usually be ignored. Additionally, the time delay is weather and climate dependent. So, the transition from one physiological stage to another has to be considered as a continuous process with continuous stage dependent parameter values. That is why we propose in this work, to build a simple model for the dynamics of tick populations where $n(s,t)$, the tick population density at the tick physiological parameter s and at time t, satisfies the McKendrick-Von Foerster PDE model. This approach has been used in a recent work by two of the authors of this paper, specifically in [5] for the case of age-structured infectious disease models, where $u(x,t)$, the density of individuals of age x at time t replaces the tick-life density $n(s,t)$. In both situations, the same McKendrick-Von Foerster model is being used, with mainly the physiological parameter s of the tick replacing the age variable x of the infectious disease model.

This paper is divided as follows. In Sect. 1, we describe the parameters that lead to the McKendrick Von Forester model (1). Then in Sect. 2, following a technique used in [5], the McKendrick Von Forester model (1) is put in the semi-variational form (2) which is validated in Sect. 3 using as in [5] a Petrov–Galerkin discretization based on a $\mathbb{P}_1$ finite-element approximation in s, followed by a semi-implicit numerical discretization in time. The resulting discrete model is then tested using the data from [8].

1 Model Description

The underlying parameters of the model are as follows:

1. As introduced above, $n(s,t)$ is the tick population density at the physiological parameter s and at time t. We will assume that the host population densities are fixed at a given density H.

2. $s_{egg}^{max}, s_{larvae}^{max}, s_{nymph}^{max}$ and s_{adult}^{max} are respectively the maximum of the physiological parameter achieved by the tick in each stage. Furthermore, the maximum of s is $s_{max} = s_{adult}^{max}$. Furthermore, the domain of s is divided into four consecutive stages: Eggs: $\Omega_1 = (s_{min}, s_{egg}^{max})$, Larva: $\Omega_2 = (s_{egg}^{max}, s_{larva}^{max})$, Nymph: $\Omega_3 = (s_{larva}^{max}, s_{nymph}^{max})$ and Adult: $\Omega_4 = (s_{nymph}^{max}, s_{max})$ and the whole domain of s can be written as $\Omega = \bar{\Omega}_1 \bigcup \bar{\Omega}_2 \bigcup \bar{\Omega}_3 \bigcup \bar{\Omega}_4$.
Also for $k = 1, 2, 3$, I_k is the interface between Ω_k and Ω_{k+1}. Thus, our interfaces are $I_1 = s_{egg}^{max}$, $I_2 = s_{larva}^{max}$ and $I_3 = s_{nymph}^{max}$.

3. **Mortality Rate:** $\mu(s,t)$ is the tick mortality rate which is host and density dependent [10], given in [6, 9] by $\mu(s,t) = \alpha + \beta ln(1 + \frac{n(s,t)}{H})$, where α and β are functions of the physiological parameter s.
Note that β depends on the stage and temperature, thus

$$\beta = \begin{cases} \text{Linear decreasing function of temperature} & \text{if} \quad s \leq s_{egg}^{max} \\ \text{Constant} & \text{otherwise} \end{cases}$$

In our model, we consider α and β as piecewise continuous functions over Ω. In other words, α and β are continuous functions within each stage Ω_i for $i = 1, 2, 3, 4$ where for any interface I, α_{I^-} and β_{I^-} are the values of the parameters to the left of the interface I and α_{I^+} and β_{I^+}, the values to the right of this interface. In a continuous case, $\alpha_{I^-} = \alpha_{I^+}$ and $\beta_{I^-} = \beta_{I^+}$.

4. **Reproduction Rate:** $K(n(s,t))$ is the reproduction rate (egg-laying) of ticks. For any interface I, denote by K_{I^-} the value of the parameter to the left of the interface I and by K_{I^+} the value of the parameter to the right of this interface. Similarly, for a continuous case, $K_{I^-} = K_{I^+}$.

5. **Seasonality and Somatic Growth or Transition Rate:** Given $d(s)$, the maxima difference function that depends on the physiological parameter s:

$$d(s) = \begin{cases} s_{egg}^{max} & \text{if} \quad s_{min} \leq s \leq s_{egg}^{max} \\ s_{larva}^{max} - s_{egg}^{max} & \text{if} \quad s_{egg}^{max} < s \leq s_{larva}^{max} \\ s_{nymph}^{max} - s_{larva}^{max} & \text{if} \quad s_{larva}^{max} < s \leq s_{nymph}^{max} \\ s_{max} - s_{nymph}^{max} & \text{if} \quad s_{nymph}^{max} < s \leq s_{max} \end{cases}$$

In [8], S. Randoph shows that, for R. appendiculatus in the laboratory or the field on Burindi, the somatic growth rate or interstadial development rate $g(s,t)$ depends on the mean day temperature T and is given by:

$$g(s,T) = \frac{d}{ae^{-bT}},$$

where d is the maxima difference and a and b are considered to be functions of the physiological parameter s that are piecewise continuous over the whole domain Ω. Thus, the somatic growth rate is a piecewise continuous function over Ω such that for any interface I, $g_{I^-}(T)$ is the value of the parameter to the left of I and $g_{I^+}(T)$ is the value of the parameter to the right of I. For a continuous case, $g_{I^-}(T) = g_{I^+}(T)$ for any temperature T.

The relationship between temperature and time can be realistically represented by a sinusoidal curve for each half of the day. Thus, the curve will take the forms

$$T = \begin{cases} (\frac{T_1+T_2}{2}) + (\frac{T_2-T_1}{2})sin(2\pi t - \frac{\pi}{2}), & 0 \le t \le \frac{1}{2} \\ (\frac{T_2+T_3}{2}) + (\frac{T_2-T_3}{2})sin(2\pi t - \frac{\pi}{2}), & \frac{1}{2} \le t \le 1 \end{cases}$$

where T_1 is the minimum temperature at the beginning of the day at $t = 0$, T_2 is the maximum temperature achieved at the midday when $t = \frac{1}{2}$ and T_3 is the new minimum at the end of the day when $t = 1$. The period between t_1 and t_2 represents the proportion of the day during which the temperature is above the threshold T_r and development takes place. The tick population density varies satisfying the following model for all $t \in [0, \mathscr{T}]$ and $s \in [s_{min}, s_{max}]$ given by

$$\begin{cases} \dfrac{\partial n(s,t)}{\partial t} + \dfrac{\partial}{\partial s}[g(s,t)n(s,t)] = -\mu(n(s,t))n(s,t) \\ \\ g(s_{min},t)n(s_{min},t) = \displaystyle\int_{s_{min}}^{s_{max}} K(n(s,t))n(s,t)\mathrm{d}s \\ \\ n(s,0) = n_0(s) \end{cases} \tag{1}$$

Let $\mathscr{N}$ be the maximum tick population density so that $n(s,t) \le \mathscr{N}, \quad \forall s \in [s_{min}, s_{max}]$ and $t \in [0, \mathscr{T}]$.

2 Variational Formulation of the Tick Life Cycle Model

In order to get the variational formulation of the tick life cycle model, we seek a solution $n(s,t)$ such that $n(.,t) \in H^1(s_{min}, s_{max})$ and $n_t(.,t) \in L^2(s_{min}, s_{max})$ $\forall\, t \in [0, \mathscr{T}]$. Let $\Gamma = \{\phi(s) \in H^1(s_{min}, s_{max})\}$ be the set of test functions and let $\phi(s) \in \Gamma$. Multiplying the first equation in the tick life cycle model by $\phi(s)$ and integrating from s_{min} to s_{max}, we get

$$< n_t, \phi > + < (gn)_s, \phi > = - < \mu n, \phi >$$

Using the definition of the mortality rate, $\mu(n(s,t)) = \alpha + \beta ln(1 + \frac{n(s,t)}{H})$ followed by integration by parts, noting the discontinuities of g on the interfaces, and using $g(s_{min}, t)n(s_{min}, t) =< K, n >$, we get the term $< (g(n))_s, \phi >$ to be equal to:

$$gn(s_{max},t)\phi(s_{max}) - < K, n > \phi(s_{min}) + \sum_{k=1}^{3}(g_{I_k^-}(t) - g_{I_k^+}(t))n(I_k,t)\phi(I_k) - < gn, \phi_s > .$$

Thus, by substitution, the variational formulation for this model becomes:

$$< n_t, \phi > + \mathscr{B}(n,\phi) =< G(n), \phi >, \tag{2}$$

where:

$$\mathscr{B}(n,\phi) =< \alpha n, \phi > - < gn, \phi_s > + g(s_{max},t)n(s_{max},t)\phi(s_{max}) - < K, n > \phi(s_{min}) + \ldots$$

$$\ldots \sum_{k=1}^{3}(g_{I_k^-}(t) - g_{I_k^+}(t))n(I_k,t)\phi(I_k)$$

and

$$< G(n), \phi > = - < \beta n \ln(1 + \frac{n(s,t)}{H}), \phi > .$$

3 Validation of the Deterministic Model

Petrov–Galerkin Procedure for the Variational Formulation of the Tick Life Cycle Model

In this section, we apply the Petrov–Galerkin procedure in order to obtain a solution of the variational formulation given by

$$< u_t, \phi > + \mathscr{B}(u,\phi) =< G(u), \phi > \quad \forall \phi \in \Gamma \tag{E}$$

where $\mathscr{B}(u,\phi) =< \alpha u, \phi > - < gu, \phi_s > + g(s_{max},t)u(s_{max},t)\phi(s_{max}) - < K, u > \phi(s_{min}) + \sum_{k=1}^{3}(g_{I_k^-}(t) - g_{I_k^+}(t))u(I_k,t)\phi(I_k)$

and $< G(u), \phi > = - < \beta u \ln(1 + \frac{u}{H}), \phi >$. Since in our case the set of test functions $\Gamma = H^1(s_{min}, s_{max})$, the problem reduces to finding $u(.,t) \in \Gamma$ such that problem (E) is satisfied.
At this stage, we apply the Petrov–Galerkin procedure on (E) based on

- Dividing the interval $[s_{min}, s_{max}]$ and $[t_0, \mathscr{T}]$ into N and M equal intervals with length h and τ respectively, to get

$$\bar{\Omega}_{h,\tau} = \{(s_i, t_j), s_i = s_{min} + i.h,\ t_j = t_0 + j.\tau,\ 0 \le i \le N \quad \text{and} \quad 0 \le j \le M\}$$

- Denoting by N_k the position of the interface I_k for $k = 1, 2, 3$. In other words, $s_{N_k} = I_k$, for $k = 1, 2, 3$.
- Simplifying the notation of the values of the parameters at the nodes s_i, $\alpha(s_i)$, $\beta(s_i)$, $K(s_i)$ and $g(s_i)$, by α_i, β_i, K_i, and g_i respectively. For any interface I_k at the position N_k, denote by $\alpha_{N_k{}^-}$, $\beta_{N_k{}^-}$, $K_{N_k{}^-}$ and $g_{N_k{}^-}$ the value of the parameters to the left of the interface I_k and $\alpha_{N_k{}^+}$, $\beta_{N_k{}^+}$, $K_{N_k{}^+}$ and $g_{N_k{}^+}$ the value of the parameters to the right of the interface I_k.
- Using finite element functions on the domain $[s_{min}, s_{max}]$, namely,

$$\phi_1(s) = \begin{cases} \frac{s_2 - s}{s_2 - s_1}, & s \in [s_1, s_2] \\ 0, & \text{otherwise} \end{cases}, \quad \phi_N(s) = \begin{cases} \frac{s - s_{N-1}}{s_N - s_{N-1}}, & s \in [s_{N-1}, s_N] \\ 0, & \text{otherwise} \end{cases}$$

$$\text{and } \phi_{i=2,\ldots,N-1}(s) = \begin{cases} \frac{s - s_{i-1}}{s_i - s_{i-1}}, & s \in [s_{i-1}, s_i] \\ \frac{s_{i+1} - s}{s_{i+1} - s_i}, & s \in [s_i, s_{i+1}] \\ 0, & \text{otherwise} \end{cases}$$

For any node s_i, denote the two parts of the finite element function ϕ_i, $i = 2, 3, ..., N-1$, defined on $[s_{i-1}, s_{i+1}]$ by

$$\phi_i{}^-(s) = \frac{s - s_{i-1}}{s_i - s_{i-1}}, \quad s \in [s_{i-1}, s_i] \quad \text{and} \quad \phi_i{}^+(s) = \frac{s_{i+1} - s}{s_{i+1} - s_i}, \quad s \in [s_i, s_{i+1}]$$

- Taking $\Gamma_N = span\{\phi_1, \phi_2, ..., \phi_N\}$.
- Choosing $\Gamma = \bigcup_{N \ge 1} \Gamma_N$ which implies that for every $u \in \Gamma$, there exists at least one $u_N \in \Gamma_N$ such that $\lim_{N \to \infty} \|u - u_N\| = 0$.

Therefore, under Petrov–Galerkin procedure, our problem reduces to finding a sequence $\{u_N\} \in \Gamma_N$ of the form $u_N(x, t) = \sum_{i=1}^N c_i(t)\phi_i(s)$ that satisfies $< (u_N)_t, \phi > + \mathscr{B}(u_N, \phi) = < G(u_N), \phi >, \quad \forall \phi \in \Gamma_N$. Thus, $\forall \phi \in \Gamma_N$, this is equivalent to $< \sum_{i=1}^N c_i'(t)\phi_i(s), \phi > + \mathscr{B}(\sum_{i=1}^N c_i(t)\phi_i(s), \phi) = - < \sum_{i=1}^N \beta_i c_i(t)\phi_i(s) ln(1 + \frac{\sum_{i=1}^N c_i(t)\phi_i(s)}{H}), \phi >$. By the bilinearity of $\mathscr{B}(u, \phi)$ and the inner product, we get a linear system of N differential equations in N unknowns, i.e., $\forall j = 1, 2, ..., N$, $\sum_{i=1}^N < \phi_i, \phi_j > c_i' + \sum_{i=1}^N [\mathscr{B}(\phi_i, \phi_j) + < \beta_i \phi_i ln(1 + \frac{\sum_{i=1}^N c_i(t)\phi_i(s)}{H}), \phi_j >]$ $c_i = 0$. This can be written in matrix form as $M \frac{dc}{dt} = -F(c)$, where $M = \{< \phi_i, \phi_j >, 1 \le i, j \le N\} \in \mathbb{R}^{N,N}$,

$F(c) = \{\sum_{i=1}^N [\mathscr{B}(\phi_i, \phi_j) + < \beta_i \phi_i ln(1 + \frac{\sum_{i=1}^N c_i(t)\phi_i(s)}{H}), \phi_j >] c_i,\ 1 \le j \le N\}$

$\in \mathbb{R}^N$, and $c = \begin{pmatrix} c_1(t) \\ c_2(t) \\ . \\ . \\ c_N(t) \end{pmatrix} \in \mathbb{R}^N$.

Write $F(c)$ as $F(c) = [-M_\alpha + A_g(t) + \Gamma_0(t) - S(c)]c$, where for $l = ln(1 + \frac{\sum_{i=1}^N c_i \phi_i}{H})$, we can write

$$M_\alpha = \begin{pmatrix} (\frac{s_2-s_1}{3})\alpha_1 & (\frac{s_2-s_1}{6})\alpha_1 & 0 & 0 & \dots & 0 \\ (\frac{s_2-s_1}{6})\alpha_2 & (\frac{s_3-s_1}{3})\alpha_2 & (\frac{s_3-s_2}{6})\alpha_2 & 0 & \dots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & \dots & (\frac{s_{N_k}-s_{N_k-1}}{6})\alpha_{N_k-} & (\frac{s_{N_k}-s_{N_k-1}}{3}\alpha_{N_k-}) + (\frac{s_{N_k+1}-s_{N_k}}{3}\alpha_{N_k+}) & (\frac{s_{N_k+1}-s_{N_k}}{6})\alpha_{N_k+} \dots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & 0 & \dots & 0 & (\frac{s_N-s_{N-1}}{6})\alpha_N & (\frac{s_N-s_{N-1}}{3})\alpha_N \end{pmatrix}$$

$$S(c) = \begin{pmatrix} \beta_1 < l\phi_1, \phi_1 > & \beta_1 < l\phi_1, \phi_2 > & 0 & \dots & 0 \\ \beta_2 < l\phi_2, \phi_1 > & \beta_2 < l\phi_2, \phi_2 > & \beta_2 < l\phi_2, \phi_3 > & 0 & \dots 0 \\ \cdot & & & & \\ \cdot & & & & \\ \dots & \beta_{N_k^-} < l\phi_{\widetilde{N}_k}^-, \phi_{N_k-1}^+ > & \beta_{N_k^-} < l\phi_{\widetilde{N}_k}^-, \phi_{\widetilde{N}_k}^- > + \beta_{N_k^+} < l\phi_{\widetilde{N}_k}^+, \phi_{\widetilde{N}_k}^- > & \beta_{N_k^+} < l\phi_{N_k}^+, \phi_{\widetilde{N}_k+1}^- > & \dots 0 \\ 0 & \dots & 0 & \beta_N < l\phi_N, \phi_{N-1} > & \beta_N < l\phi_N, \phi_N > \end{pmatrix}$$

By the midpoint rule, we have for $i \neq N_k$ for $k = 1, 2, 3$,

$$\beta_i < \phi_i ln(1 + \frac{\sum_{i=1}^N c_i\phi_i}{H}), \phi_j > = \begin{cases} \beta_i(\frac{s_i - s_{i-1}}{6})ln(1 + \frac{c_{i-1}+c_i}{2H}), \text{ for } j = i-1 \\ \beta_i[(\frac{s_i - s_{i-1}}{3})ln(1 + \frac{c_{i-1}+c_i}{2H}) + (\frac{s_{i+1}-s_i}{3})ln(1 + \frac{c_i+c_{i+1}}{2H})], \text{ for } j = i \\ \beta_i(\frac{s_{i+1}-s_i}{6})ln(1 + \frac{c_i+c_{i+1}}{2H}), \text{ for } j = i+1 \\ 0, \text{ for } j \neq i-1, i, i+1 \end{cases}$$

For $i = N_k, k = 1, 2, 3$,

$$\beta_i < \phi_i ln(1 + \frac{\sum_{i=1}^N c_i\phi_i}{H}), \phi_j > = \begin{cases} \beta_{N_k^-}(\frac{s_i - s_{i-1}}{6})ln(1 + \frac{c_{i-1}+c_i}{2H}), \text{ for } j = N_k - 1 \\ \beta_{N_k^-}(\frac{s_i - s_{i-1}}{3})ln(1 + \frac{c_{i-1}+c_i}{2H}) + \beta_{N_k^+}(\frac{s_{i+1}-s_i}{3})ln(1 + \frac{c_i+c_{i+1}}{2H}), \text{ for } j = N_k \\ \beta_{N_k^+}(\frac{s_{i+1}-s_i}{6}ln(1 + \frac{c_i+c_{i+1}}{2H}), \text{ for }, j = N_k + 1 \\ 0, \text{ for } j \neq N_k - 1, N_k, N_k + 1 \end{cases}$$

$$A_g(t) = \begin{pmatrix} \frac{-g_1(t)}{2} & \frac{g_1(t)}{2} & 0 & 0 & \dots & \dots & 0 \\ \frac{-g_2(t)}{2} & 0 & \frac{g_2(t)}{2} & 0 & \dots & \dots & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \dots & \frac{-g_{N_k-}(t)}{2} & \frac{g_{N_k-}(t)}{2} - \frac{g_{N_k+}(t)}{2} & \frac{g_{N_k+}(t)}{2} & \dots & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \dots & 0 & \dots & 0 & \frac{-g_N(t)}{2} & \frac{g_N(t)}{2} \end{pmatrix}$$

$$\Gamma_0(t) = \begin{pmatrix} (\frac{s_2-s_1}{2})K_1 & 0 \dots & \dots & \dots & \dots & 0 \\ (\frac{s_3-s_1}{2})K_2 & 0 \dots & \dots & \dots & \dots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ (\frac{s_{N_1}-s_{\{N_1-1\}}}{2})K_{N_1-} & 0 \dots & g_{N_1-}(t) - g_{N_1+}(t) & \dots & 0 & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ (\frac{s_{N_2}-s_{\{N_2-1\}}}{2})K_{N_2-} & 0 \dots & 0 & g_{N_2-}(t) - g_{N_2+}(t) & 0 & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ (\frac{s_{N_3}-s_{\{N_3-1\}}}{2})K_{N_3-} & 0 \dots & 0 & \dots & g_{N_3-}(t) - g_{N_3+}(t) & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ (\frac{s_N-s_{\{N-1\}}}{2})K_N & 0 \dots & \dots & \dots & 0 & g_N(t) \end{pmatrix}$$

Implicit Scheme

To obtain numerical results from the Petrov–Galerkin approximation applied to the variational formulation of the tick life-cycle model, we state once again the matrix form given by $M\frac{\mathrm{d}c}{\mathrm{d}t} = [-M_\alpha + A_g(t) + \Gamma_0(t) - S(c)]c$
In order to solve this system, we need to approximate $\frac{\mathrm{d}c}{\mathrm{d}t}$ by the forward difference formula, namely, $c'(t_j) = \frac{c(t_{j+1})-c(t_j)}{\tau}$
Under these approximations, we arrive at the following implicit scheme

$$\begin{cases} [M + \tau(M_\alpha - A_g(t_j) + \Gamma_0(t_j) - S(c^j))]c^j = Mc^{j-1}, \text{ for all } 1 \leq j \leq M \\ Mc^0 = F, \text{ where } F = \{< u_0, \phi_j >, 1 \leq j \leq N\} \end{cases}$$

In this stage, we use the following:

1. Data available on the tick population stated in [7].
2. Matlab sparse built-in function in order to generate the above matrices.
3. Mid-point rule to evaluate the nonlinear part $S(c)$.
4. LU decompositions.

Numerical Results

We have carried out numerical experiments using a discontinuous initial population size and piecewise constant parameters α, β, and K over a duration of 5 years starting from the year 1990. The host population density H is kept constant for the three test cases with a value of $H = 1$. In order to obtain our mesh grid, we divide the physiological interval $[s_{min}, s_{max}]$ and the time interval [1990, 1995] using the step sizes $h = 0.1$ and $dt = \frac{h}{3}$ respectively.

Using a strictly decreasing somatic growth rate g as shown in Fig. 1, we conducted the first experiment for a population initially consisting strictly of 10 eggs with the absence of ticks in the other stages. Concerning the other parameters, we used

(1) **Mortality Rate:** $\mu = \alpha + \beta ln(1 + \frac{n}{H})$ with $\alpha(s) = [0,\ 0.5,\ 1,\ 2.1]$ and $\beta(s) = [0,\ 0,\ 0.3,\ 2.8]$.
(2) **Reproduction Rate,** $K(s) = [0,\ 0,\ 10,\ 30]$.

Figure 2 allows us to notice the repetitive pattern followed by the population size over a period of five years. In fact, the population density in one year will carry on in a manner similar to those in the other years with a slight change in the initial population density depending on the outcomes of the previous year. This indicates that the tick population density $n(s, t)$ pursues periodicity in the sense that $n(s, t + 1) \sim n(s, t)$ where t is measured as a fraction of a year. This periodic property allows us to predict the behavior of the tick life cycle each year with less accurate expectations of the population density because of the continuous change of real life situations from one year to another.

We notice in Fig. 3 an evidence of the development of ticks from one stage to another. Initially, the constant number of eggs induced in the population starts to decrease gradually accompanied by a gradual increase of the number of larvae.

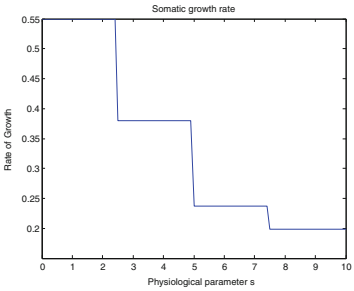


Fig. 1 Figure 1. Strictly decreasing somatic growth rate g

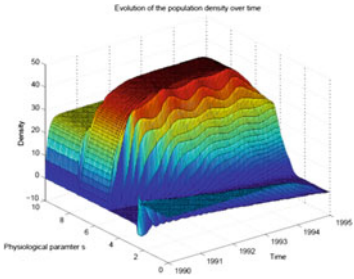


Fig. 2 Evolution of the tick population density over time for the first test case

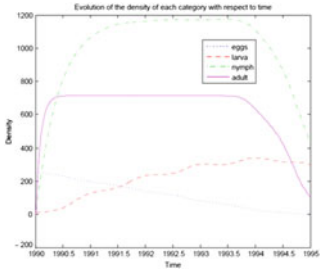


Fig. 3 Evolution of the density of each category with respect to time for the first test case

Imposing zero mortality rate for eggs allows us to say that this decrease happened when eggs developed into larvae. As the number of larvae increases, the number of nymphs increases as well to produce ticks in other stages. When nymphs reach s_{nymph}^{max}, they change into adults that are able to reproduce and give off eggs. The number of adults, as shown in the figure, remains constant over a long period of time due to the high mortality rate imposed for adults. If that wasn't the case, we would have seen an exponential growth of the number of adults due to the development of larvae and nymphs in the population.

For the second test case, the somatic growth rate presented in Fig. 4 shows a relatively low egg development in comparison with development of ticks in the other stages. We choose the same population initially consisting strictly of 10 eggs with the absence of ticks in the other stages. Concerning the other parameters, we used

(1) **Mortality Rate:** $\mu = \alpha + \beta ln(1 + \frac{n}{H})$ with $\alpha(s) = [0,\ 0,\ 0,\ 2.1]$ and $\beta(s) = [0,\ 0,\ 0,\ 2.8]$.

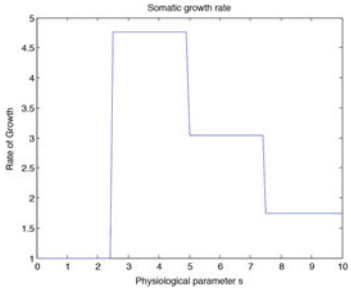


Fig. 4 Somatic growth rate g for the second test case.

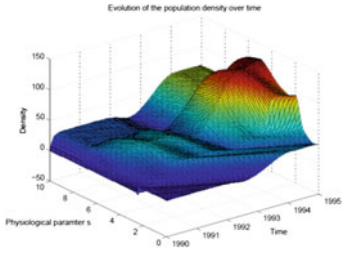


Fig. 5 Evolution of the tick population density over time for the second test case

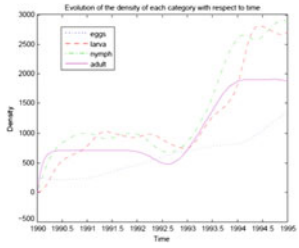


Fig. 6 Evolution of the density of each category with respect to time for the second test case

(2) **Reproduction Rate,** $K(s) = [0,\ 0,\ 0,\ 100]$.

Figure 5 shows the evolution of the tick population density over the same period of five years with less symmetry of the solution. As shown in the figure, the population density at the year 1995 shows a peak in the number of nymphs while keeping the number of adults constant over a period of 3 years.

In Fig. 6, we notice the same behavior seen in Fig. 3 in which the decrease in the number of ticks in one stage results in the increase of the number of ticks in the successive stage. However, in this case, due to the very slow development of eggs, appearance of larvae and nymphs took more time than before and eggs persisted in the population for a longer time. Development rates of larvae and nymphs allowed the population to embrace adults whose relatively high reproduction rate produced a large number of eggs as shown at the final time in the figure. Stability of the number of adults goes back to the high mortality rate imposed for this type of ticks.

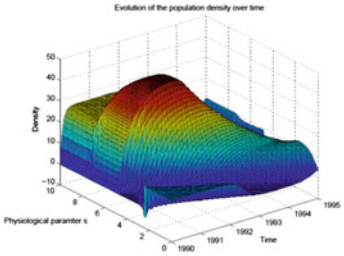


Fig. 7 Evolution of the tick population density over time for the third test case in which $g(s, t) = 1$ for all s and t

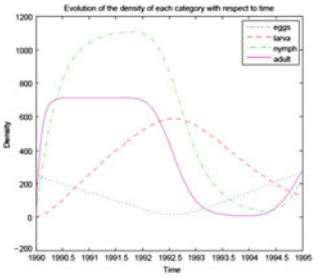


Fig. 8 Evolution of the density of each category with respect to time for the third case in which $g(s, t) = 1$ for all s and t

For the last test case, take a constant somatic growth rate $g(s, t) = 1$ for all $s \in [s_{min}, s_{max}]$ and $t \in [1990, 1995]$. As for other parameters, we keep them the same as in the second test case.

Figure 7 shows again the periodicity of the population density $n(s, t)$ in which we can predict the behavior of the solution over any year. As we can see, Fig. 8 is quite similar to Fig. 3 in the sense that the number of nymphs in the third case remains constant over a range of time quite smaller than that obtained in the first case. Also, the stability of the number of adults result from the relatively high mortality rate that prevents the development of larvae and nymphs from producing adults that persist in the population.

4 Conclusion and Future Tasks

In this paper, we have presented a simple and general finite element methodology to solve the population dynamics. Sample numerical results to validate the scheme are also presented. As for a future work, we intend to adapt the numerical method developed for the tick life cycle model to study the borne-tick interactions represented by correlated ODE-PDE equations in an attempt to test the impact of climate change on the transmission of the tick disease. On the other hand, we wish to obtain extensive additional data on tick populations interacting with hosts, in view of carrying out simultaneously statistical analysis together with our deterministic PDE model.

References

1. Bouattour, A., Darghouth, M.A., Ben Miled, L.: Cattle infestation by Hyalomma ticks and prevalence of Theileria in H. detritum species in Tunisia. Vet. Parasitol. **65**, 233–245 (1996)
2. Ghosh, M., Pugliese, A.: Seasonal population dynamics of ticks, and its influence on infection transmission: a semi-discrete approach. Bull. Math. Biol. **66**(6), 1659–1684 (2004)
3. Hudson, P.J., Dobson, A.P., Cattadori, I.M., Newborn, D., Haydon, D.T., Shaw, D.J., Benton, T.G., Grenfell, B.T.: Trophic interactions and population growth rates: describing patterns and identifying mechanisms. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **357**(1425), 1259–1271 (2002)
4. Mwambi, H.G.: Ticks and tick-borne diseases in Africa: a disease transmission model. IMA J. Math. Appl. Med. Biol. **19**(4), 275–292 (2002)
5. Nassif, N.R., Sheaib, D.: On spectral methods for scalar aged-structured population models. In: Brock, F., Saleeb, E. (eds.) Analysis and Computational Mathematics, pp. 110–123 (2009)
6. Norman, R., Bowers, R.G., Begon, M., Hudson, P.J.: Persistence of tick-borne virus in the presence of multiple host species: tick reservoirs and parasite mediated competition. J. Theor. Biol. **200**(1), 111–118 (1999)
7. Randolph, S.E.: Abiotic and biotic determinants of the seasonal dynamics of the tick Rhipicephalus appendiculatus in South Africa. Med. Vet. Entomol. **11**(1), 25–37 (1997)
8. Randolph, S.E.: Tick ecology: processes and patterns behind the epidemiological risk posed by ixodid ticks as vectors. Parasitology **129**(Suppl), S37–S65 (2004)
9. Randolph, S.E.: Dynamics of tick-borne disease systems: minor role of recent climate change. Rev. Sci. Tech. **27**(2), 367–381 (2008)
10. Rosà, R., Pugliese, A., Norman, R., Hudson, P.J.: Thresholds for disease persistence in models for tick-borne infections including non-viraemic transmission, extended feeding and tick aggregation. J. Theor. Biol. **224**(3), 359–376 (2003)
11. Rosà, R., Pugliese, A.: Effects of tick population dynamics and host densities on the persistence of tick-borne infections. Math. Biosci. **208**(1), 216–240 (2007)

Long-Term Value Creation in Mergers and Acquisitions: Contribution to the Debate

Julio Navío-Marco and Marta Solórzano-García

Abstract In recent years multiple empirical works have been undertaken to analyze the effect of mergers and acquisitions (M&A) on corporate performance, in order to effectively confirm whether M&A are investment projects that are able to create value for the shareholders. Nowadays, the analysis of the M&A value creation is still subjected to a profound debate, especially when studying the implication of time (short-term versus long-term value creation measurement) and the methodology to evaluate the value creation in the long run. In this chapter we propose a comprehensiv e review and in-depth analysis about these open debates, we evaluate the validity of the different methodologies based in calculating the stock abnormal returns provoked by the operation, and we include, as additional contribution, a short study of the implications of this theoretical discussion in a concrete example of sectoral M&A in the digital era, to illustrate the debate.

Keywords M&A · Value creation · Long-term · Short-term · Abnormal returns.

It was during the 20th century when the processes of corporate merger, acquisition and spinoff experienced their fastest growth and when they started to be the subject of in-depth, systematic studies from practically all the perspectives [36]. A review of the current state of the theoretical and empirical knowledge of M&A will reveal that the debate about their value creation evaluation still continues and there is still a long way to go in the analysis of their potential for value generation, and in the investigation of their impact on the different economic sectors.

In the next sections we propose an in-depth analysis about this open debate, especially studying the implication of time (short time versus long-term value creation

J. Navío-Marco (✉) · M. Solórzano-García
Universidad Nacional de Educación a Distancia, Madrid, Spain
e-mail: jnavio@cee.uned.es

M. Solórzano-García
e-mail: msolorzano@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_14

measurement) and the methodology to evaluate the value creation. Finally we include, to illustrate the analysis, a short study of the implications of this theoretical discussion in a concrete example of sectoral M&A in the telecommunication market.

1 Long-Term Versus Short-Term Value Creation

The need to measure the market reaction to different economic and financial events creates the need to have sufficiently rigorous analytical tools to be able to defend the validity of the results. The tools to be used and their ability to offer reliable results depend on the time horizon being considered. In this section we will see the reason for this differentiation between the short and long term as time horizons and its implications for quantification of value creation.

The study of the characteristics of securities markets, their variations and the operations carried out on these markets, including mergers and acquisitions, has in recent years been very clearly associated with an attempt to quantify their evolution over time. In this respect, two major groups can be distinguished [41]:

- Short-term market studies: they use the event study technique, which consists of quantifying the significant abnormal movements that caused the occurrence of an event (merger or acquisition) with a specific variable, which is the return on the shares.
- Long-term studies: these analyze the firm performance after the M&A and over several years, based on actual information available through the accounting books and markets.

According with [14], an event study tries to analyze the price performance at the time the event occurs and on the days before and after, in order to determine whether the prices have been affected by the event under study (in this case a M&A). The basis for such a study is therefore to estimate what return the market could expect to obtain on the day of the event, if the latter had not occurred, and compare that with what really occurred on that day and the days before and after.

In these studies, the first element that usually must be determined is the event whose impact on the prices of securities is to be studied and its date of occurrence or, otherwise, a window of days. The period over which the possible impact of the event has to be analyzed also has to be determined.

We are thus talking about an analysis that is too focused on the specific fluctuations around the date of the event, which are not always clear, that assumes a "perfect market" since that would rapidly reflect the effects of the event and it does not seem to concern itself with structural considerations or delve more deeply into the sector in which the firms are operating.

These results only reflect the market reaction in a very limited period of time, by way of fluctuations around the date of the event. It, therefore, does not appear that value creation is broadly addressed from the structural and sectoral perspective. In

this respect, some reasons why short-term movements of stock market prices do not really reflect solid value creation (or destruction) for the acquirer and its industry could include the following:

1. It is an accepted fact that short-term studies are typically limited to merger or acquisition announcements around the date of the announcement and do not address an actual situation of merger or acquisition with an effective date.
2. Short-term value changes may reflect ephemeral, speculative moves. In this respect, the merger or acquisition may affect not only the acquirer but also its competitive position, the situation of the rest of the industry and its rivals and even the likelihood of other competitors being acquired [3, 27, 47].
3. A short-term analysis window may not pick up all the effects on stock markets. There have been cases in which the shareholders of acquiring firms systematically lose value in a 3- to 5-year period after the acquisition [2, 47].
4. If an analysis of the performance of mergers and acquisitions is supported by the study of the short-term returns, this means considering that the investors fully understand the determining factors of a successful acquisition and have sufficient information to quite accurately predict how the process of integration is going to affect the future cash flows of the acquiring firm. This assumption is not likely to occur [44]. As [25] p. 129 say, "all value creation takes place after the acquisition".

2 Long-Term Analysis of M&A: The Methodological Debate

The long-term study of mergers and acquisitions is strongly conditioned by a methodological problem. The heated controversy among researchers regarding the methods for assessing long-term value creation or destruction in M&A is still to a certain extent unresolved. Therefore, to better understand the long-term perspective of mergers and acquisitions, we must first understand how to study these operations, establishing scenario in which to conduct any analysis.

We observe three major phases or eras in the research of long-term post-acquisition returns in the acquiring firms since the preliminary studies in the 1970s.

Reference [1] initially distinguished two main phases: the initial phase would include the early works of the 1980s and early 1990s, and the second phase the later, more advanced methodological research of the 1990s. However, we consider that the research has continued in the methodological sphere and the more recent studies can be included in a third phase. This third phase would begin in the late 1990s and run to the present. Thus, we would distinguish the more current methodological findings from those included in the second phase [31].

In the first phase, which primarily includes the early work between the 1970 and 1980s, the long-term analysis of returns was subordinate to short-term event studies and played a complementary role, as it was not the focal point of the research. The original long-term studies included [34, 35]. Since the appearance of abnormal

returns can indicate a contradiction with the "efficient market" assumption, the interest in observing their performance increased in the 1980s.

Reference [22] could be cited as a point of inflection in the long-term analysis of post-merger results, giving rise to what we call the second phase. They used benchmark portfolios with different factors to resolve mean-variance inefficiencies in traditional comparisons of a single factor, concluding that the findings of previous studies that indicate poor returns after the acquisition owe more to errors in the benchmark portfolios than to problems in the price at the time of the acquisition, but their results are not significant. Therefore, a more advanced methodology was introduced in this second phase in an attempt to calculate abnormal returns, and models emerged that considered explanatory factors of the returns such as size, risk and the ratio between the book value of the shareholders' equity and the market value (book-to-market ratio). It is common in the financial analysis literature [37] to list the size and the shareholders' equity book value to market value ratio as corporate features linked to a systematic risk factor. Consequently, firms to which the market assigns relatively poor expectations of wealth creation for the shareholders would show a high book-to-market ratio, meaning they would be penalized with a high capital cost, i.e. the expected return demanded to invest in them is sufficiently high to offset the tolerated risk.

There was also a more in-depth focus on the comparison with benchmark portfolios, with the relevant methodological contributions of [8, 18].

From that time to the late 1990s, the studies focused on the analysis, comparison and improvement of the estimation methods that had emerged in the previous phase:

1. BHAR, buy-and-hold abnormal returns.
2. CAR, cumulative abnormal returns.
3. CTAR, Calendar-time portfolio approach

This would be what we call phase 3, which has lasted to the present and includes new contributions, including those of [11, 16, 24, 29, 32, 33, 38, 46].

As we have indicated, there are three fundamental methodologies for analyzing these returns. We will proceed to describe them below.

Buy-and-Hold Abnormal Returns (BHAR)

This long-term abnormal return calculation method consists of compounding the short-term returns (on a monthly basis in most work) to obtain the return corresponding to the time horizon or window to be studied, based on a strategy of buying and holding during that period. Reference [43] was the first one to use this type of strategy for a long-term purchase analysis. This is thus a measure of the return that would result from investing in the securities involved in a merger or acquisition and selling them at the end of a certain time horizon, as compared to investing in certain benchmark securities.

The monthly return from the calendar month following the event to the end of the considered horizon ($s + \tau$) is estimated. In keeping with the strategy of buy-and-hold returns, the performance for a security (firm) i in a certain time horizon t, would be calculated according to the following expression:

$$\mathrm{BHR}_{i\tau} = \left[\prod_{t=s}^{s+\tau}(1 + R_{it})\right] - 1$$

where s is the calendar month of the event and R_{it} is the return of firm i in month t.

This M&A performance calculated for the sample firms is compared to a benchmark performance, thus obtaining the abnormal return (BHAR), where the cross-sectional sample mean of this abnormal return is the estimator used to measure the abnormal performance (related to the M&A solely) that could occur after the merger or acquisition.

$$\begin{aligned}\mathrm{BHAR}_{i\tau} &= \mathrm{BHR}_{i\tau} - \mathrm{BHR}_{\mathrm{CONTROL},\,\tau}\\ \overline{\mathrm{BHAR}}_{\tau} &= \sum_{i=1}^{N} w_i \cdot \mathrm{BHAR}_{i\tau}\end{aligned}$$

where N is the number of events in the sample and w_i is the weight assigned to firm i. The null hypothesis to be confirmed would be that the cross-sectional mean abnormal return is equal to zero for the sample of N firms.

Reference [8] defended the use of this method, first of all because, compared to the CAR methodology, the cumulative abnormal return is a biased estimator of the BHARs, and secondly because, even if the inference regarding cumulative abnormal returns is correct, the BHARs measure "with precision" the experience of the investor, who buys a security and holds it in portfolio during a certain period of time.

Cumulative Returns (CAR)

This method consists of calculating the excess returns with respect to a benchmark index or to the theoretical returns obtained from a certain model:

$$\overline{AR}_t = \sum_{i=1}^{N} w_i \cdot AR_{it}$$

and adding the calculated mean abnormal returns (daily or monthly) to obtain the cumulative abnormal returns (CAR).

$$CAR\tau = \sum_{t=1}^{\tau} \overline{AR}_t$$

It is then confirmed whether the mean abnormal return in each of the months that form the study time horizon is significantly different from zero.

The work of [8] demonstrating that the CAR are biased estimators of BHARs seriously undermines the reliability of using this methodology.

Calendar-Time Portfolios

This long-term return analysis methodology, used for the first time by [26, 35], consists of constructing a portfolio that each calendar month is composed of all the firms that in the τ preceding months have experienced a specific event (merger or acquisition in this case), where τ refers to the length of the event study period. The portfolio is modified every month to eliminate the firms that reach the end of the analysis period of τ months and to add firms that have undergone a merger or acquisition in the preceding month. For month t, the performance of the calendar-time portfolio is calculated as mean (or weighted mean) of the return of the sample firms that have experienced the event in the twelve, eighteen, twenty-four or thirty-six preceding months, depending on the considered time horizon.

With the obtained portfolio returns, the excess returns of the constructed portfolios are calculated for each calendar month with respect to the risk-free interest rate. Based on these excess returns, a regression is estimated with the three-factor model of [18].

The three-factor model model sustains that the expected returns of a portfolio in excess of the risk-free rate are explained by the sensitivity of its performance to three factors:

1. The excess returns with respect to a broad market portfolio or market index
2. The difference between the returns of a small enterprise share portfolio and the returns of a large enterprise share portfolio
3. The difference between the returns of a portfolio of shares with high book value versus market value, and the returns of a portfolio of shares with low book value versus market value.

The estimation of the regression model intercept is a measure of the average monthly abnormal return of the portfolio, which would be zero under the null hypothesis of absence of anomalous performance. The model is defined in the following expression:

$$R_{pt} - R_{ft} = a_p + b_p \left(R_{mt} - R_{ft}\right) + s_p SMB_t + h_p HML_t + e_{pt}$$

where R_{pt} is the return in the calendar month t of the portfolio of mergers and acquisitions made in the τ preceding months, R_{ft} is the risk-free interest rate, R_{mt} is the market portfolio return, SMB_t is the difference between the returns of portfolios composed of small and large enterprises (*small minus big*), and HML_t is the difference between returns of portfolios formed by enterprises with high and low book-to-market ratios (*high minus low*), as indicated by [18]. With these ratios, the aim is to construct two portfolios with real assets that can replicate the two non-observable risk factors and that, to the greatest possible extent, are orthogonal to each other.

If the observed abnormal returns are due to differences in risk, size and book-to-market ratio, then the estimation of the intercept (a_p) of the [18] should not be statistically different from zero.

2.1 Long-Term Mergers and Acquisitions: Main Empirical Findings

We will now review the main empirical findings obtained in the various studies that have quantitatively analyzed M&A with this long-term time horizon. This review should necessarily begin by presenting the analysis of [1], who analyzed the data, methodology and results of 22 studies of the long-term return on stock market prices of firms that made mergers and acquisitions. Table 1 includes the studies by those authors who obtained statistically significant results.

The review of [1] covers up to the end of the 20th century. Therefore, new contributions by the research community from 2000 to the present must be compiled. In Table 2 below, we include the leading studies with significant long-term results from the beginning of the century to present.

It can be observed that there is an overwhelming majority of studies, regardless of the methodology used, showing long-term negative abnormal returns in M&A, although there are examples of opposite results.

It can also be observed that most of the work analyzes horizontal samples from several sectors. It is quite unusual to find specific work on mergers and acquisitions in specific sectors, and even more in the long term. Therefore, the work of [31] in the automobile sector seems especially relevant; they find positive abnormal returns in the short run, but negative in the long run. In other industries, [15] the banking sector, which is the most active market in terms of acquisitions, yields significant positive returns for the acquirer in a three-year period. In the insurance sector, [10] detect significant long-term returns for a three-year period. In telecommunications, only [20] have focused on a statistical study of long-term mergers and acquisitions; on calculating the cumulative abnormal returns, they observe that they show significantly negative values for the years following the merger.

The evidence rejects the equality of mean abnormal returns across industries at significant levels. While a number of industries such as petroleum and natural gas, insurance and machinery, experienced significantly positive abnormal performance, others like business services and medical equipment have demonstrated significantly negative long-term returns. Consistent with prior research findings, the results suggest around zero long-term performance for acquisitions in the banking industry. Reference [48].

2.2 Statistical Considerations Regarding the Possible Methodologies

Once we have reviewed the different methodologies and the research literature that have quantitatively analyzed M&A in the long-term run, we proceed to present a comparative analysis of the different methods' characteristics, to better determine their explanatory capacity and robustness, for the purpose of determining which one

Table 1 Studies of long-term mergers and acquisitions up to 2000 with significant results

Study	Sample	Methodology	Date	Results
Asquith [6]	196 successful completed offers and 87 unsuccessful between 1962-1976	Beta control portfolio	Completed	–7.2% cummulative abnormal returns for the successful proposals and –9.6% for unsuccessful
Malatesta [34]	256 USA acquirers between 1969 and 1974	Beta control portfolio	Announcement date	Abnormal returns of –5.4% for +1 to +6 months. Abnormal returns of –2. 2% in a period of +7 a +12 months
Franks and Harris [21]	1858 operations between 1975 and 1984	Three methods	24 months after closing	Values between –0.126 a 0.048
Agrawal, Jaffe and Manoflker [2]	670 M&As between 1966 and 1987	Size and ratio accounting value/market value	Completed	Meanful abnormal returns of –1.026 for 60 months
Anderson and Mandelker [5]	452 UK companies between 1984 and 1992	Six methods	Completed	Meanful abnormal returns between –0.0931 and –0.0956 for 60 months
Gregory [23]	452 UK companies between 1984 and 1992	Six methods	Completed	Meanful values depending on method between –1.182 and –0.18 for 24 months
Loughran and Vijh [32]	947 companies (788 operations and 135 proposals)	BHAR	Completed	Meanful abnormal returns (–25%) for 60 months; unmeanful results for the offers
Rau and Vermaelen [42]	2823 operations and 316 offers between 1980 and 1991	Control portfolio with size and ratio accounting value/market value	Completed	Meanful abnormal returns of –4% for 36 months; meanful abnormal returns +8.56% for the offers
Mitchell and Stafford (1998) (*)	2767 acquisitions between 1961 and 1993	Several methods: BHAR, CTAR (with Fama and French regression)	Completed	Meanful results for equally weighted portfolios

Source Adapted by authors from [1].
(*) Early version of [38].

Table 2 Studies of long-term mergers and acquisitions from 2000 to present with significant results

Study	Sample	Methodology	Date	Results
Mitchell and Stafford [38]	2767 acquisitions between 1961 and 1993	BHAR and CTAR (with Fama & French regression)	Completed	Meanful results for equally weighted portfolios
Moeller, Schlinge-mann and Stultz [39]	12023 US acquisitions between 1980 and 2001	BHAR;CTAR	Completed	BHAR: meanful abnormal return of –16.02% for 36 months; CTAR: no meanful abnormal returns
Sudarsanam and Mahate [45]	519 UK acquirers between 1983 and 1995	Size and ratio accounting value/market value	Completed	Meanful abnormal returns between –8.71% and –21.89%
Delong [15]	54 acquirers from banking between 1983 and 1995	BHAR	Completed	Meanful abnormal returns of 1.1% for 3-years
Gregory and McCorriston [24]	197 UK acquisitions in USA, 97 of UK in UE, and 39 of UK in other countries	BHAR (and CAR for short-term)	Completed	Meanful abnormal returns of –9.36% and –27.1% between +3 and +5 years in US; no meanful results for Europa
Conn, Cosh, Guest and Hugues [12]	131 public companies, private	CTAR; BHAR	Completed	Acquierers of public companies lose –19.78% on average in 36 months
Alexandridis, Antoniou and Petmezas [4]	179 acquirers	CTAR and CAMP	Completed	Meanful abnormal loses between –0.55% and 1.02% for both models
Kyriazis [30]	86 operations in greek market	Fama and French	Completed	2% monthly losses during 3 years after acquisition
Laab and Schiereck [31]	230 M&A between 1981 and 2007 en automotion	CTAR; BHAR	Announcement date	Meanful value destruction between 16% and 20% for three years
Craninckx and Huyghebaert [13]	267 public operations and 336 private operations between 1997 and 2006	CAR and BHAR	Completed	Define a fail ratio for the acquisitions of almost 50%
Baker, Dutta, Saadi and Zhu [7]	1066 acquisitions	CAR	Announcement date	Operating performance falls meanfully for acquirers that previously showed higher operating performance

Source Prepared by authors

is more reliable. In this respect, a series of statistical considerations, that will help to throw some light on the advantages and drawbacks of each methodology, is presented in this section.

Poor Model Specification

One problem that affects the definition of long-term abnormal return assessment models is the poor model specification. Of the possible results obtained, it is advisable to determine how much is really an abnormal return and how much is simply the result of a poor model specification. However, the problem of poor model specification is, in fact, a implicit difficulty in any statistical modeling exercise.

In this respect, in their studies on measuring long-term performance, [29] concluded that the BHAR and CAR methods are conceptually erroneous and/or lead to statistical contrasts with rejection rates exceeding the nominal values that are usually used, as they tend to find a positive or negative abnormal return when in fact it does not exist.

Furthermore, [38] indicate that BHARs give a false impression of the adjustment speed. Abnormal returns appear to persist over a long period, although they may only occur for a much shorter time. Moreover, they argue that BHARs increase within larger time horizons. As this horizon cannot be derived from theory but is chosen arbitrarily by the researchers, "it is impossible to infer from the analysis of BHARs over different holding periods how long an abnormal return actually persists." [16].

On the other hand, [17] considers that the BHAR methodology is not very appropriate because the errors caused by incorrect specification of the model generating expected returns are compounded by calculation of the long-term returns.

But it is actually the calendar-time portfolio method (CTAR) that is criticized the most for poor model specification, as is also affected by the criticisms already voiced against the three-factor model of [18] that it uses. If the factors introduced by Fama and French (size, book-to-market ratio) are not able to fully explain the returns, then the rejection or acceptance of the null hypothesis may be questionable because it is not known if it corresponds to the abnormal returns or to the poor model specification.

To evaluate a potential poor model specification problem in CTAR, [38] conduct an experiment in order to have a clearer view of what effect the deficiencies of the three-factor model have, and they break down the intercept into two components: on one hand the expected abnormal performance in accordance with the sample characteristics (size, book-to-market ratio and M&A frequency over time), and on the other the amount of abnormal performance attributable to other sources, including the M&A itself. The expected intercept, contingent on the sample composition, is estimated as the mean of one thousand time series regressions of as many random samples composed of firms similar to the event firms (i.e. the M&A) but that have not had one. They use this calculation to determine the distribution mean under the null hypothesis, not to measure the dispersion. Furthermore, they select samples with a size, frequency and book-to-market ratio similar to those of the event benchmark portfolio and calculate a new statistic t using the expected intercept and the original intercept.

Their results indicate that a third of the monthly mean abnormal returns estimated for the samples results from the poor model specification and not from the event. In any case, as [19], p. 70, emphasize, "although the amount and value of the associated t decrease, the poor performance of acquisitions is still statistically significant after making the mentioned adjustment", thus accepting the relevance of the abnormal results obtained.

Selection of the Benchmark in the Return's calculation

As seen in all the methodologies described above, a key point is to appropriately select the benchmark that serves as the basis for comparison of the cumulative returns. In other words, we ask ourselves what would be the benchmark return against which our firm or sample shows abnormal returns. An inadequate benchmark selection clearly leads to obtain inaccurate or false abnormal returns.

The possible alternatives would be:

1. Use a benchmark firm that serves as the basis for comparison
2. Use a market index
3. Use a benchmark portfolio

The use of a single firm as benchmark of the long-term returns is problematic and inappropriate. The length of the study intervals hinders the selection of the benchmark firm and can introduce a survivorship bias. Furthermore, the definition of the decision criteria for selecting this firm can also be difficult. Other disadvantages are the possible emergence of idiosyncratic risk factors (i.e. inherent in the selected firm) and the possibility of increases as the time horizon increases. This means that the firm specific factors may be dominant, "reducing the calculation of the abnormal returns to a lottery" [16], p.34.

Many authors use a market index as benchmark, but [33] recommend using a well constructed benchmark portfolio.

Reference [38] criticize BHARs for being far from the null distribution and claim that using a carefully structured benchmark portfolio may have little impact with this method. Moreover, [8] acknowledge that the BHAR results are affected by the periodic readjustment of the benchmark portfolio.

Biases

Barber and Lyon demonstrate that cumulative abnormal returns are biased estimators of BHARs, especially when the market index is used as benchmark. This has been one of the main arguments that have discouraged the use of CAR.

As for the BHAR methodology, [8, 29] identify three biases that affect the use of statistical contracts when benchmarks such as market indexes or size-constructed portfolios are used as benchmark. These biases appear as a result of:

1. Listing of new securities
2. Readjustment of the benchmark portfolio

3. Asymmetry of the multi-period abnormal returns (the distribution appears as deviated to the right and not centered on zero)

and they make statistical inferences more difficult [19].

Reference [8] propose a solution matching sample firms to control firms of similar sizes and book-to-market ratios and specify this method to mitigate the three identified biases, but they recommend further research to progress with their solution. On the other hand, to correct the asymmetry of the long-term abnormal returns obtained by this method, [33] recommend, in order to obtain well specified contrasts, to use resampling techniques with replacement (bootstrapping).

Reference [8] identify another potential bias that could affect the different methodologies: depending on how the benchmark is calculated, a survivorship bias can arise from considering strong benchmark firms that survive and endure in the market. The impact of this possible bias is dismissed by these authors in [9], where they present evidence that the survivorship bias in the databases they use does not significantly affect the estimation.

Calendar-time portfolios can also present problems of bias; the fact that different sectors have different parameter estimations leads to biased estimations when the focus of the sample portfolio composition switches from one sector to another.

Cross-Sectional Dependence

Reference [33] proposed a robust BHAR estimation approach to the three biases mentioned above, postulating appropriate benchmark portfolios that could mitigate these undesired effects; but this method do not solve the biggest problem of the BHAR method, and is the main reason for arguing against this methodology: it has not been able to avoid the cross-sectional dependence.

Reference [33] define three potential sources of dependence in the abnormal returns that affect the BHAR method: superimposition of the mean returns, clustering by calendar time and clustering by industry.

The mean returns become superimposed when the first merger or acquisition of the firm is followed by a second one during the period in which the annual returns of the first are being measured. For example, estimating the BHARs three years after two events that are only one year apart would cause a superimposition of twenty-four month returns. It is obvious that the dependence problem cannot arise when a firm is associated with only one event.

Clustering by calendar and clustering by industry would also introduce dependence in the form of cross-correlation. According to [16], there is evidence that samples are clustered by time and industry. Therefore, the returns in one period of firms in the same industry would overlap. Intuitively, the problem of dependence increases with the sample size and the time horizon to which the BHAR method is applied. Under equivalent conditions, the possibility of superimposition increases with the number of observations in a sample period. Likewise, the longer the estimated period, the greater will the possibility be of another event occurring in the same period.

The problems associated with the cross-sectional correlation of BHARs in non-random samples are hard to correct. Reference [11, 33, 38] present methodologies that try to reduce the poor specification of the contrasts. However, these methods suffer from two major problems: they are only well specified for random samples, and the use of bootstrapping is not able to solve the problem of dependence, as indicated by [38].

On the contrary, the calendar-date portfolio methodology manages to avoid the problem of cross-sectional dependence. In this method, when the portfolio is constructed, the variance in each of the periods automatically incorporates the cross-sectional correlation of the individual returns of the sample firms. According to [16], establishing a portfolio of firms that have experienced the event in a certain month is the most robust method to control cross-dependence. For instance, if a portfolio is created with the firms that have undergone a merger or acquisition in the thirty-six preceding months, it can be readjusted on a monthly basis to eliminate the mergers and acquisitions that are more than thirty-six months old and add new ones. This approach is a monthly portfolio readjustment. Consequently, the returns of the established portfolio account for the cross-correlation because there is no inference based on standard cross-sectional errors.

Reference [38] provide evidence to defend the calendar-time portfolio approach versus the compounding of returns, claiming it is more a powerful way to detect the poor performance of the first methodology versus the cross-sectional dependence of the individual abnormal returns found in the second method. In this respect, the calendar-time portfolio approach shows a clear advantage over BHARs.

Heteroskedasticity

Unfortunately, not everything is advantageous in the calendar-time portfolio approach. The procedure of creating monthly portfolios automatically eliminates the problem of cross-sectional dependence, but it introduces an additional complexity; because the calendar-time portfolios contain a variable number of firms and they are continually modified from one month to the next, a change occurs in the regression residuals and the possibility of heteroskedasticity arises.

To solve this issue, Mitchell and Stafford require that at least ten firms be included in each benchmark portfolio; "we mitigate the heteroskedasticity problem substantially by requiring at least 10 firms in the event portfolio at each point in time, which accounts for the majority of the diversification effect of the portfolio residual variance." Reference [38], p.316.

Other authors try to overcome this potential limitation by modeling the conditional variance with the standard GARCH(1,1) model, but the results do not differ substantially [16].

In spite of the aforementioned drawbacks, the calendar-time portfolio approach, as compared to the previously analyzed methodologies (BHARs and CARs), offers the tremendous advantage that, with the construction of the portfolio, the variance in each of the periods automatically incorporates the cross-sectional correlation of the individual returns of the sample firms. The use of large samples and the careful construction of benchmark portfolios can partially mitigate the negative effects of the

BHAR methodology, according to citeLyo and [38], but it cannot solve the serious problem of cross-sectional dependence.

Therefore, and even more importantly, because the serious statistical problems commented above cannot be avoided, we clearly recommend the calendar-time portfolio approach because, as [38], p. 296, say: "Since our objective is to reliably measure abnormal returns, it is imperative that the methodology allows for reliable statistical inferences."

At any rate, it is common for some authors to choose to complete the calendar-date portfolio calculations with the results provided by the other methodologies. New improvements are also being introduced into the BHAR approach [19], to try to resolve the aforementioned serious drawbacks, which will open up the field even more for financial and econometric researchers.

Consequently, we also recommend to include in any research the results obtained with the other mentioned methodologies for purposes of analysis completeness and robustness because, as [17] had already anticipated, "it is recommendable to compare the results obtained from different methods".

3 Ilustrating the Theoretical Debate with an Empirical Analysis of M&A

To conclude this study of long-term value creation evaluation, we include the results of the analysis developed by the authors in [40] studying the long-term value creation in the telecommunication sector from 1995–2010. This study is especially relevant here to illustrate the different results obtained using the three described methodologies and the debate between long-term and short term approaches.

Based on this econometric analysis, this research found clear evidence of value destruction (2000–2010) in the long run, when the sector is analysed as a whole, and evidence of value creation when the companies involved use the same language.

The study analyses 4337 M&A made between operators, and prepares 720 benchmark portfolios for comparison. Although the chosen methodology was the calendar-time portfolio approach, these authors include, for purposes of method completeness and robustness, the values obtained using the other methodologies and calculate also the calendar-time cumulative abnormal returns (Table 3), which are defined as the mean abnormal return calculated every month for each firm, subtracting from the monthly portfolio returns of each firm the expected portfolio return [38].

$$CTAR_t = R_{pt} - E(R_{pt})$$

In all the cases, with all the methodologies, there is evidence of the progressive "negativization" of the sample mean. In other words, as we move towards long-term time horizons, we evolve towards value destruction by mergers and acquisitions in telecommunications. This result reinforces the concerns about short term studies that

Table 3 Comparison of results of the different methodologies

		CTAR	BHAR	CAR
3 months	*n*	282	280	282
	mean	0.0030221	0.0100773	–0.0017163
	std. deviation	0.175769	0.175067	0.207724
	t-statistic	0.288729	0.963205	–0.138751
	p-value	0.773	0.3363	0.8897
6 months	*n*	281	283	281
	mean	–0.0008868	–0.0056857	–0.0108269
	std. deviation	0.228348	0.235013	0.2755
	t-statistic	–0.0651029	–0.406991	–0.658773
	p-value	0.9481	0.6843	0.5106
12 months	*n*	263	262	263
	mean	–0.0245416	–0.004326	–0.0549688
	std. deviation	–1.24107	–0.231845	–2.28785
	t-statistic	0.2157	0.8168	0.02294
	p-value	0.2157	0.8168	0.02294
24 months	*n*	250	232	250
	mean	–0.053437	–0.0370553	–0.113869
	std. deviation	0.420127	0.358437	0.54822
	t-statistic	–2.01109	–1.57464	–3.28414
	p-value	0.04539	0.1167	0.00117
36 months	*n*	195	191	194
	mean	–0.129255	–0.019915	–0.274171
	std. deviation	1.01001	0.661473	1.24105
	t-statistic	–1.78707	–0.416088	–3.07705
	p-value	0.07549	0.6778	0.002395

Source [40]

can conduct to inaccurate conclusions such as a positive impression of value creation in the M&A; longer time horizon can demonstrate value destruction giving a more complete and accurate view on the effects of the operation. This can be observed graphically in Fig. 1.

Also in all the cases, with the different time horizons, when calendar-time portfolios are used as benchmark the values are closest to zero (smallest values in absolute value), which seems consistent with the perfect market hypothesis because it is not possible to expect large abnormal returns. In this respect, the values with BHAR and CAR are greater than with the calendar-time portfolio methodology showing consistency with their indicated theoretical disadvantages. The same sign is obtained with the three methods in each time frame. With all the methodologies, we observe that after six months, value would already be destroyed as all the methods present negative signs in the sample mean, although the results are significant with longer horizons according to the methods.

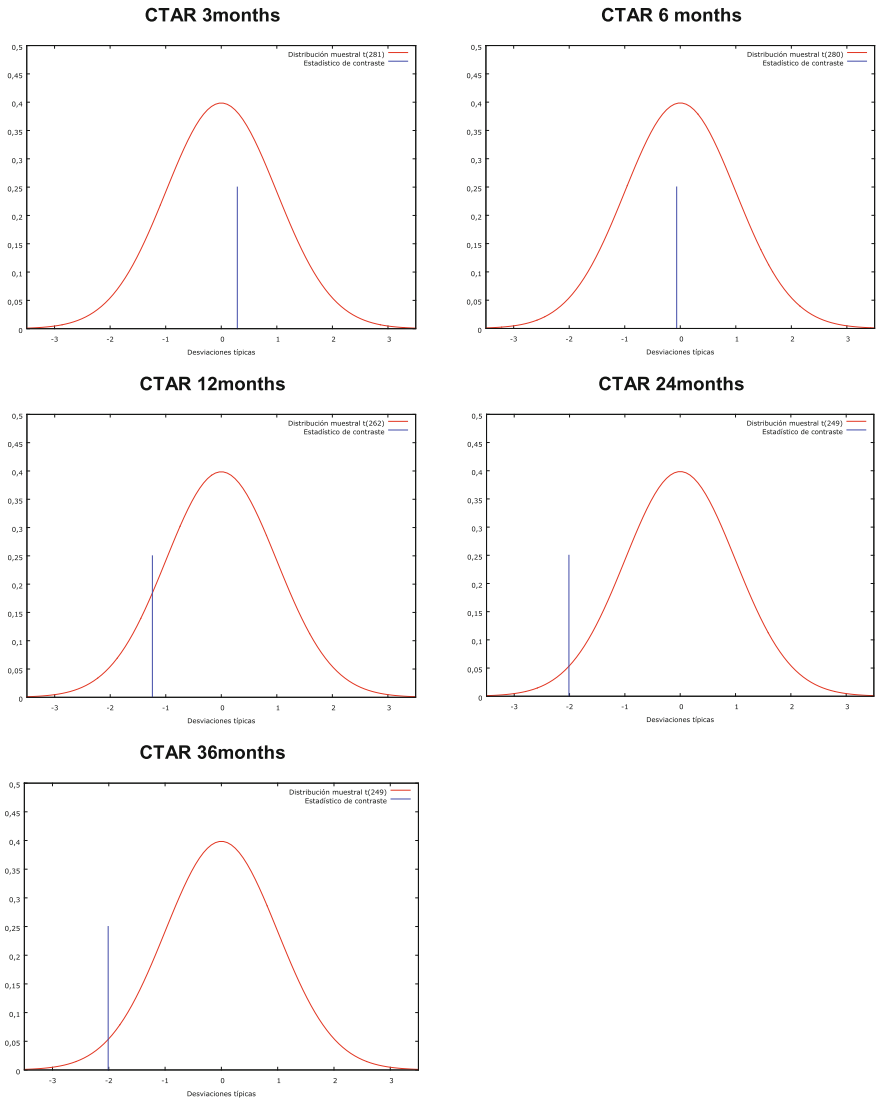


Fig. 1 Graphic representation of CTAR evolution. *Source:* Prepared by authors based in [40]

References

1. Agrawal, A., Jaffe, J.F.: The post-merger performance puzzle. In: Cooper, G., Gregory, A. (eds.) Advances in Mergers and Acquisitions. Elsevier, Amsterdam (2000)
2. Agrawal, A., Jaffe, J.F., Mandelker, G.N.: The Post-Merger performance of acquiring firms: a re-examination of an anomaly. J. Financ. **47**(4), 1605–1621 (1992)

3. Akhigbe, A., Madura, J.: The industy effects regarding the probability of takeovers. Financ. Rev. **34**, 1–18 (1999)
4. Alexandridis, G., Antoniou, A., Petmezas, D.: Divergence of opinion and postacquisition performance. J. Bus. Financ. Account. **34**(3), 439–460 (2007)
5. Anderson, C., Mandelker, G.: Long run return anomalies and the book-to-market effect: evidence on mergers and IPOs. Unpublished Working Paper. (1993)
6. Asquith, P.: Merger bids, uncertainty, and stockholder returns. J. Financ. Econ. **11**(1-4), 51–83 (1983)
7. Baker, H.K., Dutta, S., Saadi, S., Zhu, P.: Are good performers bad acquirers? Financ. Manag. **41**(1), 95–118 (2012)
8. Barber, B.M., Lyon, J.D.: Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. J. Financ. Econ. **43**, 341–372 (1997a)
9. Barber, B.M., Lyon, J.D.: Firm size, book-to-market ratio, and security returns: a holdout sample of financial firms. J. Financ. **52**(2), 875–883 (1997b)
10. Boubakri, N., Dionne, G., Triki, T.: Consolidation and value creation in the insurance industry: the role of governance. J. Bank. Financ. **31**(1), 56–68 (2008)
11. Brav, A.: Inference in long-horizon event studies: a bayesian approach with application to initial public offerings. J. Financ. **55**, 1979–2016 (2000)
12. Conn, R., Cosh, A., Guest, P., Hugues, A.: The impact on UK acquirers of domestic, cross-border, public and private acquisitions. J. Bus. Financ. Account. **32**(5), 815–870 (2005)
13. Craninckx, K., Huyghebaert, N.: Can stock markets predict M&A failure? a study of european transactions in the fifth takeover wave. Eur. Financ. Manag. **17**(1), 9–45 (2011)
14. Del Brío, E.B.: Descripción Metodológica de los Estudios de Evento a Corto Plazo. Serie Teknos **1**, 1–59 (2009)
15. DeLong, G.L.: The announcement effects of U.S. versus non-U.S. bank mergers: do they differ? J. Financ. Res. **26**(4), 487–500 (2003)
16. Ecker, F.: Information Risk and Long-Run Performance of Initial Public Offerings. Gabler, Wiesbaden (2008)
17. Fama, E.F.: Market efficiency, long-term returns and behavioral finance. J. Financ. Econ. **49**, 283–306 (1998)
18. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. J. Financ. Econ. **33**, 3–56 (1993)
19. Farinós, J.E., García, C.J., Ibáñez, A.M.: Problemas en la estimación y en el contraste de los rendimientos anormales a largo plazo: Estado de la cuestión. Cuadernos de Gestión **2**(2), 51–77 (2002)
20. Ferris, S.P., Park, K.: How different is the long-run performance of mergers in the telecommunications industry? Innov. Invest. Corp. Financ. **7**, 127–144 (2002)
21. Franks, J.R., Harris, R.S.: Shareholder wealth effects of corporate takeovers: the UK experience 1955–1985. J. Financ. Econ. 23(2), 225–249 (1989)
22. Franks, J., Harris, R., Titman, S.: The postmerger share-price performance of acquiring firms. J. Financ. Econ. **29**(1), 81–96 (1991)
23. Gregory, A.: An examination of the long run performance ofUKacquiring firms. J.Bus. Financ. Account. **24**, 971–1002 (1997)
24. Gregory, A., McCorriston, S.: Foreign acquisitions by UK limited companies: short and long-run performance. J. Empir. Financ. **12**, 99–125 (2005)
25. Haspeslagh, P.C., Jemison, D.B.: Managing Acquisitions. Free Press, New York (1991)
26. Jaffe, J.F.: Special information and insider trading. J. Bus. **47**, 410–428 (1974)
27. Kohers, N., Kohers, T.: Information sensitivity of high tech industries: evidence from merger announcements. Appl. Financ. Econ. **14**, 525–536 (2004)
28. Kothari, S.: Capital markets research in accounting. J. Account. Econ. **31**, 105–231 (2001)
29. Kothari, S.P., y Warner, J.B.: Measuring Long-Horizon security price performance. J. Financ. Econ. **43**, 301–339 (1997)
30. Kyriazis, D.: The long-term post acquisition performance of greek acquiring firms. Int. Res. J. Financ. Econ. **43**, 69–79 (2010)

31. Laabs, J.-P., Schiereck, D.: The long-term success of M&A in the automotive supply industry: determinants of capital market performance. J. Econ. Financ. **34**, 61–88 (2010)
32. Loughran, T., Vijh, A.M.: Do long-term shareholders benefit from corporate acquisitions? J. Financ. **52**(5), 1765–1790 (1997)
33. Lyon, J., Barber, B., Tsai, C.-H.: Improved methods for tests of long-run abnormal stock returns. J. Financ. **54**(1), 165–201 (1999)
34. Malatesta, P.H.: The wealth effect of merger activity and the objective functions of merging firms. J. Financ. Econ. **1**(4), 155–181 (1983)
35. Mandelker, G.: Risk and return: the case of merging firms. J. Financ. Econ. **1**(4), 303–335 (1974)
36. Mascareñas, J.: Fusiones y adquisiciones de empresas. McGraw-Hill, Madrid (2005)
37. Marín, J.M., Rubio, G.: Economía Financiera. Antoni Bosch, Barcelona (2001)
38. Mitchell, M.L., Stafford, E.: Managerial decisions and long-term stock price performance. J. Bus. **73**(3), 287–329 (2000)
39. Moeller, S.B., Schlingemann, F.P., Stulz, R.M.: Do shareholders of acquiring firms gain from acquisitions? (No. w9523). National Bureau of Economic Research (2003)
40. Navío-Marco, J., Solórzano-García, M., Matilla-García, M., Urueña, A.: Language as a key factor of long-term value creation in mergers and acquisitions in the telecommunications sector. Telecommun. Policy **40**, 1052–1063 (2016)
41. Palacín, M.J.: El mercado de control de empresas: el caso español. Ariel, Barcelona (1997)
42. Rau, P.R., Vermaelen, T.: Glamour, value and the post-acquisition performance of acquiring firms. J. Financ. Econ. **49**(2), 223–253 (1998)
43. Ritter, J.R.: The long-run performance of initial public offerings. J. Financ. **46**(1), 3–27 (1991)
44. Sorescu, A., Chandy, R., Prabhu, J.C.: Why some acquisitions do better than others: product capital as a driver of long-term stock returns. J. Mark. Res. **44**(1), 57–72 (2007)
45. Sudarsanam, P.S., Mahate, A.: Glamour acquirers, method of payment and postacquisition performance: the UK evidence. J. Bus. Financ. Account. **30**, 299–341 (2003)
46. Tuch, C., O'Sullivan, N.: The impact of acquisitions on firm performance: a review of the evidence. Int. J. Manag. Rev. **9**(2), 141–170 (2007)
47. Walker, M.M., Hsu, C.: Strategic objectives, industry structure, and the long-term stock price performance of acquiring and rival firms. Appl. Financ. Econ. **17**(15), 1233–1244 (2007)
48. Yaghoubi, R., Locke, S., Gibb, J. L.: Acquisition Returns: Does Industry Matter? (2012). http://ssrn.com/abstract=2022860

Cournot Duopolies with Investment in R&D: Regions of Nash Investment Equilibria

B. M. P. M. Oliveira, J. Becker Paulo and Alberto A. Pinto

Abstract We study a model of a Cournot duopoly where firms invest in R&D to reduce their production costs. Depending on the parameters, we may find regions with one, two or three Nash equilibria of the investment. Here, we study the effect of the parameters in these regions, in particular, we study the effect of the possible market saturation, the maximum relative cost reduction and the product differentiation, giving special attention to regions with multiple Nash equilibria. We observed that, in general, the competitive region, where both firms invest, is reduced as we increase the possible market saturation and the differentiation of the products and is enlarged when we increase the maximum relative cost reduction.

Keywords Nash equilibria · Cournot duopoly · Multiple equilibria · R&D investment

1 Introduction

We consider a Cournot duopoly where firms invest in R&D to reduce their production costs, as described in [3, 4, 7, 11]. This competition is modeled, as usual, by a

B. M. P. M. Oliveira (✉)
FCNA, Universidade do Porto, Portugal and LIAAD-INESC TEC, Porto, Portugal
e-mail: bmpmo@fcna.up.pt
URL: http://www.orcid.org/0000-0002-7710-4284

J. Becker Paulo
FC, Universidade do Porto, Porto, Portugal
URL: http://www.orcid.org/0000-0003-4651-3808

A. A. Pinto
FC, Universidade do Porto, Portugal and LIAAD-INESC TEC, Porto, Portugal
e-mail: aapinto1@gmail.com
URL: http://www.orcid.org/0000-0003-2953-6688

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_15

two stages game. In the first subgame, the two firms invest in R&D to reduce the respective initial production costs. In the second subgame, the two firms are under Cournot competition, with production costs equal to the reduced cost determined by the R&D investment. We use the R&D cost reduction function introduced in [7]. In [1, 5–10] we study the perfect Nash equilibria of this two stages game and the economical effects of these equilibria. The second subgame, consisting of a Cournot competition, has a unique perfect Nash equilibrium. For the first subgame, consisting of an R&D cost reduction investment program, we exhibit four different regions of Nash investment equilibria that we characterize as follows: a competitive Nash investment region C where both firms invest, a single Nash investment region S_1 for firm F_1, where just firm F_1 invests, a single Nash investment region S_2 for firm F_2, where just firm F_2 invests, and a nil Nash investment region N, where neither of the firms invest (see [7, 9]). These regions may intercept, and we observe, for some parameter values, regions with one, two or three Nash investment equilibria.

2 R& D Investments on Costs

We consider an economy with a monopolistic sector with two firms, F_1 and F_2, each one producing a differentiated good. The inverse demands p_i are linear:

$$p_i = \alpha - \beta q_i - \gamma q_j, \tag{1}$$

with parameters: *value to buyers* $\alpha > 0$, *possible market saturation in monopoly* (market saturation) $\beta > 0$ and *product differentiation* γ. We assume that $\gamma > 0$ and thus the goods are substitutes. Firm F_i may invest an ammount $v_i \geq 0$ in an R&D program $a_i : \mathbb{R}_0^+ \to [b_i, c_i]$ that reduces its production cost to

$$a_i(v_i) = c_i - \frac{\epsilon(c_i - c_L)v_i}{\lambda_i + v_i}, \tag{2}$$

with $b_i = a_i(+\infty) = c_i - \epsilon(c_i - c_L)$ and with parameters: *minimum production cost* c_L in $(0, \alpha)$, *initial production cost* c_i in $[c_L, \alpha]$, *maximum relative cost reduction* (maximum reduction) ϵ in $(0, 1)$ and *inverse of the quality of the R&D investment program* $\lambda > 0$, see [7] for further details. Here we will consider that the firms will only compete in one period of time. Furthermore, we assume that the two firms are identical except, at most, in their production costs.

The profit $\pi_i(q_i, q_j)$ of firm F_i is given by

$$\pi_i(q_i, q_j) = q_i(\alpha - \beta q_i - \gamma q_j - a_i) - v_i, \tag{3}$$

for $i, j \in \{1, 2\}$ and $i \neq j$. The Nash equilibrium output (q_1^*, q_2^*) is given by

$$q_i^* = \begin{cases} 0, \ if \ R_i \leq 0 \\ R_i, \ if \ 0 < R_i < \frac{\alpha - a_j}{\gamma} \\ \frac{\alpha - a_i}{2\beta}, \ if \ R_i \geq \frac{\alpha - a_j}{\gamma} \end{cases}, \tag{4}$$

with R_i as defined in [5].

The new production costs region can be decomposed, at most, in three disconnected economical regions characterized by the optimal output level of the firms: the *monopoly region* M_1 of firm F_1, the *duopoly region* D, and the *monopoly region* M_2 of firm F_2. For further details see, e.g., [5, 7]. The boundaries between the duopoly region D and the monopoly region M_i are l_{M_i} with $i \in \{1, 2\}$ and are presented, explicitly in [7–9].

The profit function π_i of firm F_i, is given by

$$\pi_i(a_1, a_2) = \begin{cases} \pi_{i,M_i}, & \text{if } (a_1, a_2) \in M_i \\ \pi_{i,D}, & \text{if } (a_1, a_2) \in D \\ -W_i(a_1, a_2), & \text{if } (a_1, a_2) \in M_j \end{cases},$$

with π_{i,M_i}, $\pi_{i,D}$ and W_i as defined in [5]. The best investment response function $V_i : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ of firm F_i is explicitly computed in [7]. This can be a multi-valued function.

We study the effect of the *normalized production differentiation* (normalized differentiation), $\hat{\gamma} = \gamma/\beta$, with $\hat{\gamma} = 1$ being its default value. The default values of the other parameters are $c_L = 4$, $\alpha = 10$, $\beta = 0.013$, $\epsilon = 0.2$ and $\lambda = 10$. In a neighbourhood of these values, we found that the Nash investment equilibria consists of a unique, or two, or three points, depending upon the pair of initial production costs [7]. The set of all Nash investment equilibria form the *Nash investment equilibrium set*, that can be divided in three types of regions: the *competitive Nash investment region* C, the *single Nash investment region* S_i, and the *nil Nash investment region* N. For further details see e.g. [5, 7]. The nil Nash investment region can be further decomposed into 4 regions: N^{LL}, when the production costs of both firms are low; N^{HH}, when the costs of both firms are high; N^{LH}, when the first firm has low production cost and the second has high cost; and N^{HL}, the symmetric of the previous case. Moreover, the single investment region for each firm can also be decomposed into: the single favorable region S_i^F, when the only firm that invests is the one with the lower production cost, thus enhancing its advantage; and the single recovery region S_i^R, when the only firm that invests is the one with the high production cost, thus reducing its disadvantage. The region of multiple Nash investment equilibria are the result of the interception of (at least two of) S_i; S_j and C [5, 7].

3 Nash Investment Equilibria

The regions of Nash investment equilibria depend on the initial production costs and on the other parameters. We observe that multiple Nash investment equilibria may be found in the region of high production costs. In this section, we study the effect on the regions when we change the following parameters: market saturation β, maximum reduction ϵ, and normalized production differentiation $\hat{\gamma}$. Regarding the coloring of Figs. 1, 2 and 4, the nil Nash investment region N is painted in grey if the firms are in duopoly, and dark blue or dark red if one of the firms is in monopoly. The single Nash investment regions S_1 and S_2 are colored blue and red, respectively, using a slightly lighter tone if firms are in duopoly. The competitive Nash investment region C is painted in green. The region where S_1 and S_2 intersect are colored pink, the region where S_1 and C intersect are painted in light blue and the region where S_2 and C intersect are colored yellow. The region where the regions S_1, S_2 and C intersect are colored light grey.

We begin by studying the effect of an increase in the market saturation β, representing a change in the outputs of the firms, from cases with higher quantity output ($\beta = 0.0002$) through 0.0013 (its default value) to cases with lower output ($\beta = 0.0100$), see Fig. 1. For the lower values of β we observe a large competitive region C and very small nil investment regions N^{LL}, N^{LH}, and N^{HL}. Furthermore, the single favourable region S_i^F present for both firms. We observe, for each firm, the presence of a very thin single, recovery region S_i^R. We could not observe regions of multiple Nash investment equilibria with the numeric accuracy we used. When we increase β, the competitive region is reduced and we observe an increase of the area covered by the nil investment regions N^{LL}, N^{LH}, N^{HL} and an increase of the single recovery region S_i^R for both firms. The presence the regions with multiple Nash equilibria and the nil N^{HH} region are observed in a neighborhood of the default values of the parameters. When the value of β is further increased, we observe that the nil investment regions are enlarged, eventually merging, while the single and the competitive region shrink until they disappear. The last regions to disappear are the single favourable regions inside the duopoly region. If β becomes large enough, the profit of the firms is so small that it will not be worthy to invest.

We also study the effect of the maximum relative cost reduction ϵ, see Fig. 2. For low values of ϵ, the firms do not invest, since the return from the investment is insufficient. Hence, the nil investment region occupies the parameter space we are analysing. As we increase ϵ to 0.03, we observe that the single favourable investment regions S_i^F and S_j^F appear in the duopoly regions for intermediate costs, and in the monopoly region, for costs near the boundary with the single investment regions inside the duopoly region. As we further increase the maximum reduction ϵ, the single favourable investment regions S_i^F increase and we observe the appearance of the competitive region ($\epsilon = 0.05$) and the nil investment region is split. We also

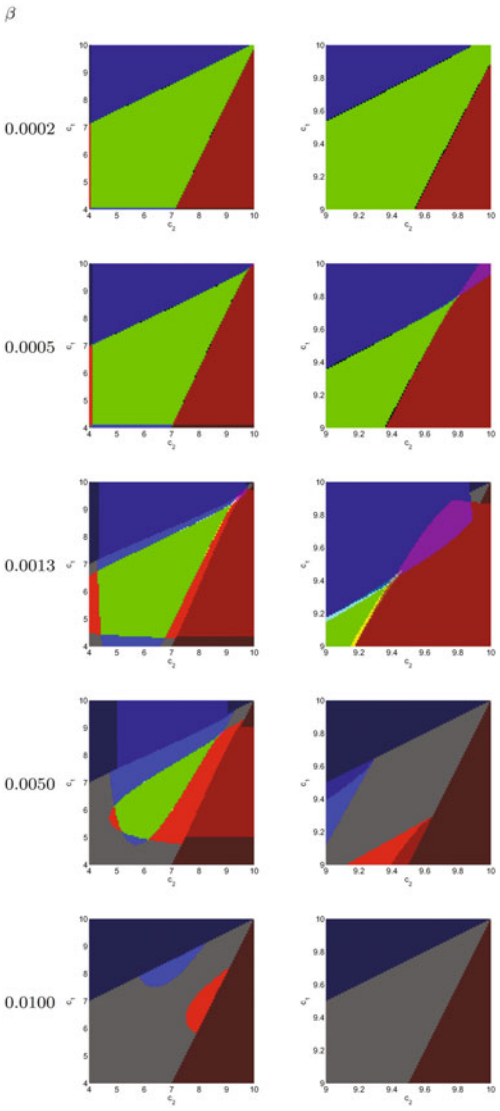


Fig. 1 Effect the market saturation of β in the regions of Nash investment equilibria. Production costs in [4, 10]. Zoom with production costs in [9, 10], $\epsilon = 0.2$ and $\hat{\gamma} = 1$. See Sect. 3 for details in coloring

observe the presence of two small single recovery regions S_i^R. We find another split in the nil investment regions, creating N^{LL}, N^{LH}, N^{HL} and N^{HH} for larger values of ϵ. This occurs as the competitive region increases, pushing the also increasing recovery regions into the boundaries. For the default value, $\epsilon = 0.20$, all these regions

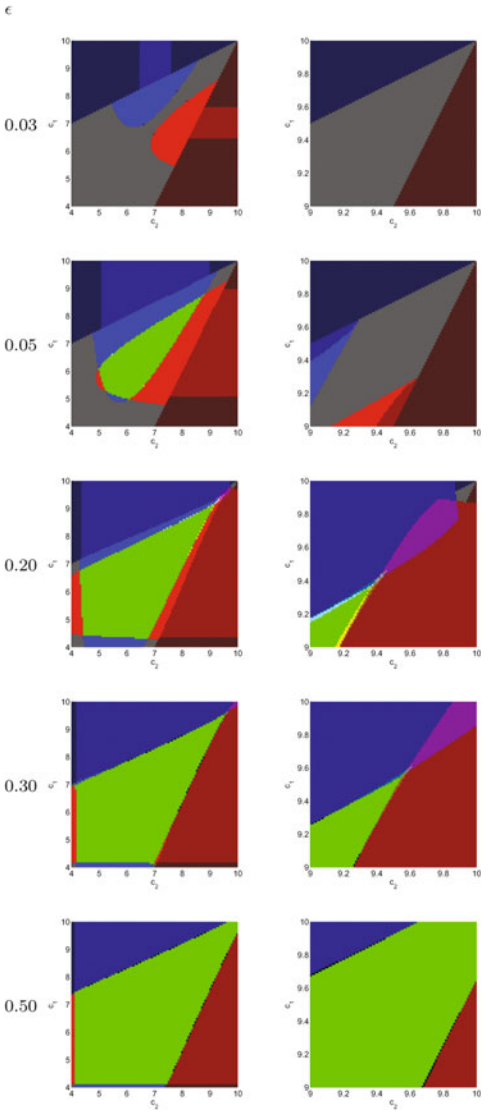


Fig. 2 Effect maximum reduction of ϵ in the regions of Nash investment equilibria. Production costs in [4, 10]. Zoom with production costs in [9, 10]. $\beta = 0.0013$ and $\hat{\gamma} = 1$. See Sect. 3 for details in coloring

are visible, together with the regions with multiple Nash equilibria. Increasing the maximum reduction ϵ to larger values (0.30 and 0.50), makes the investment more effective. This causes the shrinkage of the single investment regions in duopoly,

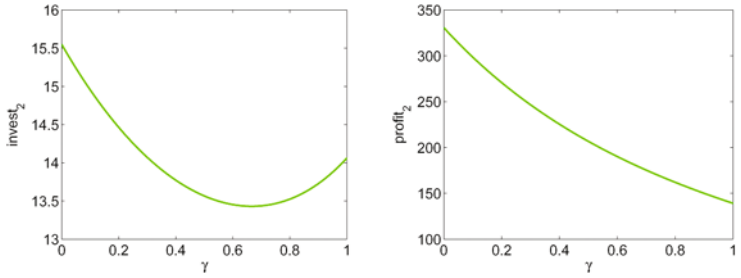


Fig. 3 Effect of the normalized product differentiation $\hat{\gamma} = \gamma/\beta$ on the investment (left) and on the profit (right). The investment decreases until $\hat{\gamma} \simeq 0.7$ and increases afterwards. The profit decreases with $\hat{\gamma}$. The initial costs for both firms are $(c_i, c_j) = (6, 6)$. All other parameters are at their default values

together with the multiple investment regions, while the competitive region increases and the monopoly regions suffer little change.

Finally, we study the effect of the the normalized product differentiation $\hat{\gamma}$, in $(0, 1]$, see Figs. 3 and 4. When we study the investment and the profits as a function of the normalized product differentiation, we observe that the investment is minimum for $\hat{\gamma} \simeq 0.7$, see Fig. 3. We also observe that the profit decreases monotonically with $\hat{\gamma}$. The decrease in the profit can be explained by the increased competition each firm suffers when $\hat{\gamma}$ is increased, thus decreasing its products selling price. Regarding the investment, for low values of $\hat{\gamma}$, close to 0 as the profit is reduced, so it is reduced the amount invested. For large values of $\hat{\gamma}$, close to 1, as $\hat{\gamma}$ increases, the competition increases and there is a more evident advantage to invest. A reduction in the production cost of one firm will allow it to increase its output, while the other firm may be forced to reduce its outputs. For our parameter values, the effect in the investment of the decrease in the profit, while not having enough advantage in the investment create the observed minimum of the investment when $\hat{\gamma} \simeq 0.7$.

We also studied how the regions change as $\hat{\gamma}$ changes, see Fig. 4. If $\hat{\gamma} = 0$, the two goods are considered to be independent and the market behaves like each firm is in monopoly. When the products are substitutes for $\hat{\gamma} > 0$ we observe that there is an interaction effect between the investment and the profits of the firms. This creates the possibility of multiple Nash investment regions, that we find when $\hat{\gamma} = 0.2$. As we increase $\hat{\gamma}$, we observe that the competitive region decreases, as all single investment region increase, in particular, the single favourable investment in monopoly. The nil investment regions have different behaviour, while N^{LL}, N^{LH} and N^{HL} increase, in particular the latter two, N^{HH} have a similar total size, although both monopoly regions inside it increase with $\hat{\gamma}$.

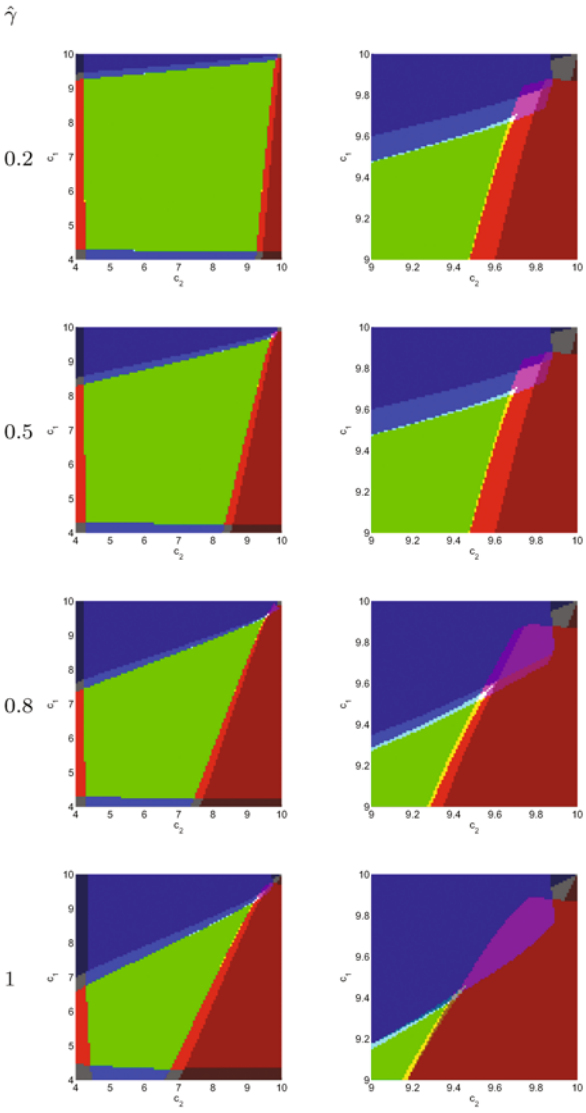


Fig. 4 Effect normalized differentiation of γ in the regions of Nash investment equilibria. Production costs in [4, 10]. Zoom with production costs in [9, 10]. $\beta = 0.0013$ and $\epsilon = 0.2$. See Sect. 3 for details in coloring

4 Conclusions

We studied the Cournot competition model with R&D programs, using the R&D investment function introduced in [7]. In this chapter, we gave special attention to the competitive, single and nil regions investment. In each region, the firms could be in duopoly or in monopoly. We observed the effect on these regions of the Nash investment equilibria when we changed the parameters: possible market saturation

in monopoly β, maximum relative cost reduction ϵ and normalized product differentiation $\hat{\gamma}$. We found, as in [1, 5, 7–10], that there are regions in the parameter space where the Nash investment equilibrium is not unique, in these regions two or three equilibria can be found. In this article we observed the persistence of the regions of each type and described how these regions change as we change the parameters β, ϵ and $\hat{\gamma}$ of the R&D programs of both firms. Overall, the competitive region decreases as we increase the market saturation β and the normalized differentiation $\hat{\gamma}$ and increases as we increase the maximum reduction ϵ. The single investment regions and the nil regions have opposite behaviour to the competitive region. Furthermore, we also observed here that the regions with multiple Nash equilibria are present for high production costs for both firms.

Acknowledgements This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project Dynamics, optimization and modelling PTDC/MAT-NAN/6890/2014 and by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

References

1. Becker, J., Ferreira, M., Oliveira, B.M.P.M., Pinto, A.A.: R&D dynamics. Discret. Contin. Dyn. Syst. 61–68 (2013). Supplement
2. Brander, J.A., Spencer, B.J.: Strategic commitment with R&D: the symmetric case. Bell J. Econ. **14**, 225–235 (1983)
3. Cournot, A.: Researches into the Mathematical Principles of the Theory of Wealth. In: Bacon, N. (ed.) Recherches sur les Principes Mathématiques de la Théorie des Richesses. Paris, 1838. English edition: Macmillan, New York (1897)
4. d'Aspremont, C., Jacquemin, A.: Cooperative and noncooperative R&D in duopoly with spillovers. Am. Econ. Rev. **78**, 1133–1137 (1988). (Erratum. In American Economic Review 80, 641-642)
5. Ferreira, M., Almeida, J.P., Oliveira, B.M.P.M., Pinto, A.A.: R&d dynamics with asymmetric efficiency. In: Cushing, J.M., Pinto, A.A., Elaydi, S., i Soler, L.A. (eds.) 18th International Conference on Difference Equations and Applications, ICDEA 2012, Springer Proceedings in Mathematics and Statistics. pp. 73–83 (2016)
6. Ferreira, M., Figueiredo, I.P., Oliveira, B.M.P.M., Pinto, A.A.: Strategic optimization in R&D investment. Optimization **61**(8), 1013–1023 (2012)
7. Ferreira, M., Oliveira, B.M.P.M., Pinto, A.A.: Patents in new technologies. J. Differ. Equ. Appl. **15**, 1135–1149 (2009)
8. Ferreira, M., Oliveira, B.M.P.M., Pinto, A.A.: R&D bankruptcy boundaries determined by patents. In: Peixoto, M.M., Pinto, A.A., Rand, D.A. (eds.) Dynamics, Games and Science II, DYNA 2008. Springer Proceedings in Mathematics. vol. 2 (26), pp. 275–300 (2011)
9. Ferreira, M., Oliveira, B.M.P.M., Pinto, A.A.: Piecewise R&D dynamics on costs. Fasciculi Mathematici **44**, 29–42 (2010)
10. Pinto, A.A., Oliveira, B., Ferreira, F.A., Ferreira, M.: Investing to survive in a duopoly model. Intelligent Engineering Systems and Computational Cybernetics. Springer, Netherlands (2008). Chap. 23
11. Singh, N., Vives, X.: Price and quantity competition in a differentiated duopoly. RAND J. Econ. **15**, 546–554 (1984)

A Stochastic Logistic Growth Model with Predation: An Overview of the Dynamics and Optimal Harvesting

S. Pinheiro

Abstract We consider a logistic growth model with predation and a stochastic perturbation given by a diffusive term with power-type coefficient. The resulting stochastic differential equation (SDE) has the particularity that the standard conditions for the existence and uniqueness of solutions of SDEs do not hold for a large subset of parameter space. Thus, we start by discussing the well posedness of the problem at hand, leading to a detailed characterization for the existence and uniqueness of solutions. We then provide criteria ensuring extinction and persistence of such population. Additionally, we list subsets of parameter space where (absolutely continuous) stationary measures for the SDE under consideration are guaranteed to exist, providing a description for the corresponding densities. We conclude with an application to the optimal management of resources. We consider a real asset such as, for instance, a farm or an aquaculture facility, devoted to the exploration of a unique culture or population whose growth follows a SDE such as described above, and look for the optimal harvesting strategy associated with such culture or population.

Keywords Population dynamics · Stochastic differential equations · Optimal control

1 Introduction

We provide an overview of the main results in the papers [27–29] and the PhD thesis [26], concerning the asymptotic dynamics and an optimal harvesting problem associated with a stochastic logistic growth model with a predation term given by a Holling type-n functional, for some integer $n \geq 2$, and a stochastic part driven by a one-dimensional standard Brownian motion with a diffusion coefficient of power-

S. Pinheiro (✉)
Dept. of Mathematics, Brooklyn College, City University of New York, New York, NY, USA
e-mail: Susana.Pinheiro95@brooklyn.cuny.edu

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_16

type, i.e. proportional to $x^{\alpha}\mathrm{d}W_t$, where x is the population size, α is some positive constant and W is a standard one-dimensional Brownian Motion.

We start by discussing the existence and uniqueness of global solutions for the Stochastic Differential Equation (SDE)

$$\mathrm{d}x(t) = \left[\rho x(t)\left(1 - \frac{x(t)}{K}\right) - \varepsilon\frac{(x(t))^{n-1}}{1 + (x(t))^{n-1}}\right]\mathrm{d}t + \sigma(x(t))^{\alpha}\mathrm{d}W_t\,, \qquad (1)$$

with positive initial condition and up to the first instant of time where such solution reaches zero. Given the setup under consideration here, the only meaningful and relevant continuation for a solution reaching zero, is for such solution to remain constant and equal to zero afterwards. There are two key difficulties to be addressed. The first one is that the coefficients of (1) do not satisfy the linear growth condition, so that solutions may explode in finite time. The second issue that needs to be addressed is that whenever $\alpha < 1$ the diffusion coefficient of (1) is not Lipschitz continuous in any neighbourhood of zero, which may lead to problems concerning uniqueness of solutions. We overcome such problems by introducing a modified locally Lipschitz condition, which turns out to be particularly suitable for the problem under consideration here. Moreover, we use Lyapunov functions techniques to prove that solutions of (1) do not blow up to infinity in finite time.

Previous related work includes Roberts and Saha paper [31] concerning the asymptotic behaviour of logistic epidemic models. We should also mention that Gary et al. [15] extend the classical SIS epidemic model from a deterministic framework to a stochastic one, yielding a stochastic perturbation of a logistic differential equation whose diffusion coefficient is (a quadratic polynomial) proportional to the population deterministic growth rate. In [17, 18] Jiang et al. consider a randomized logistic equation with coefficients given by periodic functions and a diffusive coefficient depending linearly on the population size. In [16] Ji et al. consider a stochastic logistic equation, but with no predation term and the additional assumption that the parameter α is restricted to the interval $(1, 3/2)$. In [8, 9] Braumann considers a large family of stochastic differential equations modelling the growth of populations subjected to harvesting in a randomly fluctuating environments is considered. Conditions for non-extinction and for the existence of stationary distributions are provided for these models. However, the SDE under consideration here does not fit into the assumptions used there. Similar results can be found in [3, 14, 24] and the references therein, for specific density-dependent natural growth functions and harvesting policies.

In what concerns the Optimal Harvesting Problem, early developments of the subject were related with deterministic population dynamics models, for both discrete-time and continuous-time systems (see, e.g. [10] and references therein for further details on the subject). Earlier approaches to this topic consisted in finding harvesting strategies maximizing sustained yields [3, 24], under the working assumption that a (absolutely continuous) stationary distribution exists for the population size, ignoring the risk of population extinction. The combined influence of extinction from demographic and environmental stochasticity, as well as from harvesting strategies,

was studied in [20], where the criteria used to find the optimal harvesting strategies was the maximization of the cumulative harvest subject to some prescribed risk of extinction. Some more recent works on the topic of optimal harvesting have focused mainly on the use of stochastic optimal control techniques to maximize the expected total discounted amount of the harvested population under the assumption that the population size evolves according to some form of the stochastic logistic growth model [1, 21].

The approach followed in [26, 29] uses dynamic programming techniques to find harvesting policies that jointly maximize the utility derived from continuously harvesting part of the population over a finite interval of time and the utility obtained from reaching the final instant in that interval with the largest possible population size. Thus, the problem surveyed here combines the point of view of maximizing utility from harvesting, while also aiming at long-term population preservation. For the sake of completeness we mention that Dynamic Programming was first developed by Bellman in the 1950s [4–7], and further extended by Florentin [12, 13] and Kushner [19]. The key goal in the dynamic programming methodology is to obtain a backwards recursive relation for the value function associated with a given optimal control problem. If additional regularity conditions are satisfied, such recursive relation can be written as a boundary value problem associated with a second order partial differential equation known as the Hamilton–Jacobi–Bellman (HJB) equation. Rather complete discussions of this subject may be found in the monographs [11, 32]. It should be remarked that difficulties arise when trying to implement dynamic programming techniques to address the optimal harvesting problem associated with the stochastic logistic growth model described earlier. The main problems are caused by the fact that the linear growth condition does not hold for the drift part of our stochastic logistic growth SDE and the fact that the diffusion coefficient is not always Lipschitz continuous. Nevertheless, under a weaker set of assumptions, we obtain in [26, 29] a dynamic programming principle and the corresponding HJB equation for the stochastic optimal control problem we are interested in. We then apply these results to find the optimal harvesting strategies in feedback form and to provide a couple of qualitative properties for the value function, namely in what concerns its monotonicity with respect to the state variable and some relevant model parameters.

2 On a Stochastic Logistic Growth Model with Predation

Throughout this section we will consider the stochastic logistic growth model (1). We start by noticing that (1) does not satisfy the standard assumptions for existence and uniqueness of solutions. However, under a weak set of assumptions, we were able to ensure that solutions with a positive initial condition exist and are unique up to the first instant of time at which zero is reached. Moreover, we provide criteria for population extinction, persistence and for the existence of a stationary measure. Furthermore, we provide a detailed characterization for the asymptotic stationary measure density in the former case. The contents of this section are part of [27, 28].

2.1 The Stochastic Model

Let $\left(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P}\right)$ be a complete probability space with a filtration $\{\mathscr{F}_t\}_{t\geq 0}$ satisfying the usual conditions, i.e. $\{\mathscr{F}_t\}_{t\geq 0}$ is increasing and right continuous while $\mathscr{F}_0$ contains all $\mathbb{P}$-null sets. Let $\{W_t\}_{t\geq 0}$ be a standard one-dimensional Brownian motion defined on the filtered probability space $\left(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P}\right)$. We will consider stochastic perturbations of the logistic growth model with predation term of the form

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + g(x(t))\mathrm{d}W_t \,, \tag{2}$$

where

$$f(x) = \rho x\left(1 - \frac{x}{K}\right) - \varepsilon\frac{x^{n-1}}{1+x^{n-1}} \,, \tag{3}$$

and

$$g(x) = \sigma x^{\alpha} \,, \tag{4}$$

for some positive real parameters $\rho, K, \varepsilon, \sigma, \alpha$ and integer $n \geq 2$.

The additive stochastic term driven by the Brownian motion W models fluctuations in the size of the population caused by a number of external factors such as, for instance, changes in weather and climate, as well as the influence of diseases and competition with other species. In the following sections we will provide conditions that guarantee the existence and uniqueness of relevant solutions of (2) and we will discuss the asymptotic behaviour of such solutions by providing criteria ensuring extinction of the population or persistence of the population, as well as the existence of an absolutely continuous stationary measure.

2.2 Existence of Global Solutions

In this section we state the existence of a global non-negative solution of (2) for every positive initial condition. There are two main difficulties to be addressed. The first one is that the linear growth condition does not hold for the drift part of (2), which could cause the solution to blow up in finite time. The second one is that for values of α smaller than one, the diffusive part of (2) is not Lipschitz in a neighbourhood of 0, which may disturb the existence of local solutions.

We say that a function $f : \mathbb{R} \to \mathbb{R}$ is *locally Lipschitz in* $\mathbb{R}^+$ if for every integer $k \geq 1$ there exists $h_k > 0$ such that for every $x, y \in \left[\frac{1}{k}, k\right]$ we have

$$|f(x) - f(y)| \leq h_k\,|x - y| \,.$$

We remark that the definition given above is not the standard definition for locally Lipschitz functions, but rather an adjusted version addressing the issue caused by (4) not being Lipschitz on any neighbourhood of zero if $\alpha < 1$.

We state next an abstract result implying the existence of local solutions of (2). Its proof extends that of Theorem 3.15 of [22] to the setup under consideration here. See [26, 27] for further details.

Proposition 1 *Fix $T > 0$ and let $f, g : \mathbb{R} \to \mathbb{R}$ be two locally Lipschitz functions in $\mathbb{R}^+$. Then, the stochastic differential equation*

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + g(x(t))\mathrm{d}W_t$$

has a unique maximal local solution in $\mathbb{R}^+$, $\{x(t)\}_{0 \le t \le \sigma_\infty}$, for every positive initial condition x_0.

Thus, as a consequence of Proposition 1 we obtain that for every $T > 0$ and for every $\alpha > 0$, the stochastic differential equation (2) admits a unique maximal local solution in $\mathbb{R}^+$, $x(t)$, defined on $[0, \sigma_\infty)$. If $\sigma_\infty = T$, then $x(t)$ is finite and strictly positive for all $t \in [0, T]$. Otherwise, $x(t)$ becomes zero or blows up to infinity at the instant $\sigma_\infty < T$. We will see next that the solutions of (2) never blow up to infinity. Moreover, if $\alpha \ge 1$ the solutions remain positive for every $t \ge 0$. However, if $\alpha < 1$ the solutions of (2) may reach zero in finite time. Nevertheless, from the point of view of the problem under consideration here, the only meaningful continuation for a solution reaching zero in finite time is to have it constant and equal to zero from that instant of time onwards. More precisely, let τ_0 be the stopping time defined as

$$\tau_0 = \inf\{t \ge 0 : x(t) = 0\}\,,$$

and let $x(t)$ be such that

$$x(t) = \begin{cases} x(0) + \int_0^{t \wedge \tau_0} f(x(s)) + \int_0^{t \wedge \tau_0} g(x(s))\mathrm{d}s & \text{if } t < \tau_0 \\ 0 & \text{if } t \ge \tau_0 \end{cases}\,. \tag{5}$$

From now on we restrict our attention to the set $\mathscr{M}$ of solutions of (2) of the form (5), i.e. with the property that if there exists $t^* > 0$ such that $x(t^*) = 0$, then $x(t) = 0$ for every $t > t^*$. We obtain the following result.

Theorem 1 *For any given positive initial condition $x(0) = x_0$ and any $\alpha > 0$, the SDE has a unique global non-negative solution $x(t)$ in $\mathscr{M}$. Moreover, if $\alpha \ge 1$ the solution is strictly positive.*

The following result follows as a consequence through the use of a stochastic dominance argument.

Corollary 1 *Consider the family of stochastic differential equations of the form*

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + \sigma(x(t))^\alpha \mathrm{d}W_t \quad t \ge 0\,, \tag{6}$$

where σ and α are positive constants. Assume that f is locally Lipschitz in $\mathbb{R}^+$ and that the following additional conditions hold:

(i) $f(0) = 0$
(ii) *there exist* $A, B > 0$ *such that* $f(x) \leq Ax(B - x)$ *for every* $x > 0$.

Then, for any positive initial condition $x(0) = x_0$*, the SDE* (6) *has a unique non-negative solution in* $\mathscr{M}$.

For a proof of the two statements above see [26, 27].

2.3 Criteria for Population Extinction

We will now give criteria ensuring extinction of populations whose size evolves according to the logistic stochastic model with predation term (1). More precisely, we provide conditions under which extinction of the population occurs, respectively, with positive probability and full probability. Additionally, we provide extra conditions under which the population becomes extinct exponentially fast with full probability. More importantly, in the special case where $\alpha < 1$, we prove that extinction occurs in finite time with full probability. From an intuitive point of view, population extinction with full probability when $\alpha < 1$, regardless of every other parameter values, can be explained by the combination of two factors. First, notice that in the absence of any randomness the population size tends to be below the (finite) carrying capacity of the drift term of (1). Second, the variance of the instantaneous rate of growth of the population increases with decreasing population size, becoming arbitrarily large when the population size approaches zero. Thus, since the population size can not escape to infinity, the population eventually becomes extinct in finite time whenever $\alpha < 1$ due to the large variance of its instantaneous rate of growth.

The next theorem states that for every $\alpha < 1$ the population will become extinct with probability one. For its proof see [26, 27].

Theorem 2 *If* $\alpha < 1$ *and* $x_0 > 0$*, then*

$$\mathbb{P}\{\exists t < \infty : x(t) = 0\} = 1 \,,$$

that is, $x(t)$ *reaches zero a.s.. In other words, the population goes extinct with probability one.*

We will now consider the case $\alpha \geq 1$. Very roughly, we obtain that the population becomes extinct with positive probability for sufficiently small values of the natural growth rate ρ.

Theorem 3 *Assume that one of the following conditions holds:*

- $\alpha = 1$, $n = 2$ *and* $\rho - \varepsilon < \sigma^2/2$
- $\alpha = 1$, $n > 2$ *and* $\rho < \sigma^2/2$
- $\alpha > 1$, $n = 2$ *and* $\rho < \varepsilon$.

Then, the trivial solution $x(t) = 0$ of (2) is stochastically asymptotically stable, i.e. there exists a set of positive Lebesgue measure $A \subset \mathbb{R}^+$ such that for any initial condition $x_0 \in A$, the solution of (2) through x_0 becomes extinct with positive probability.

We will now provide conditions under which the trivial solution $x(t) = 0$ is almost surely exponentially stable. As we will see below, such conditions are strongly related with the existence of a negative upper bound for the infinitesimal generator associated with the SDE (2) acting on the logarithm function $\ln x$. To fix notation, let $F : \mathbb{R}^+ \to \mathbb{R}$ denote the aforementioned infinitesimal generator:

$$F(x) = \rho\left(1 - \frac{x}{K}\right) - \varepsilon\frac{x^{n-2}}{1 + x^{n-1}} - \frac{1}{2}\sigma^2 x^{2\alpha-2}\,. \tag{7}$$

We note that, given some choice of parameters, exactly one of the following three alternative descriptions holds for F:

(i) F is strictly decreasing for every $x > 0$;
(ii) there exists $x^* > 0$ such that F is strictly increasing in $[0, x^*)$ and strictly decreasing for every $x > x^*$;
(iii) there exist $0 < x_1 < x_2$ such that F is strictly decreasing in $[0, x_1)$, strictly increasing in (x_1, x_2) and strictly decreasing for every $x > x_2$.

Moreover, we remark that each one of the three alternative behaviours above holds in a subset of parameter space with positive Lebesgue measure.

Theorem 4 *Assume that one of the conditions of Theorem 3 holds and that, additionally, the function F in (7) admits a strictly negative upper bound. Then, for any given positive initial condition, the solution of the SDE (2) obeys*

$$\limsup_{t\to\infty} \frac{1}{t}\log(x(t)) < 0 \quad a.s.\,,$$

namely, the trivial solution $x(t) = 0$ is almost surely exponentially stable. In other words, the population goes extinct with probability one, exponentially fast.

For a proof of the previous two theorems see [26, 27].

2.4 Persistence

We now shift our focus to the subject of population persistence and the existence of a (absolutely continuous) stationary measure under the dynamics of the SDE (1). Indeed, such measures play an analogue role to stable equilibria in the corresponding deterministic models. Furthermore, under the same set of conditions as for persistence, we state that a unique stationary distribution exists. Finally, we use the Forward

Kolmogorov equation to provide a rather detailed characterization for the densities of the stationary measures of (1), summarizing all the information in a "stochastic" bifurcation diagram. We remark that such bifurcation diagram is complete in the sense that the subset of parameter space where persistence of population occurs (and an absolutely continuous stationary measure exists) is the complement in parameter space of the subset where population extinction is guaranteed to occur (up to the measure zero subset corresponding to the bifurcation thresholds).

Theorem 5 *Assume that one of the following sets of conditions holds:*

- $\alpha = 1$, $n = 2$ *and* $\rho - \varepsilon > \sigma^2/2$
- $\alpha = 1$, $n > 2$ *and* $\rho > \sigma^2/2$
- $\alpha > 1$, $n = 2$ *and* $\rho > \varepsilon$
- $\alpha > 1$ *and* $n > 2$.

Then, for any given positive initial condition x_0 *the solution of the SDE* (2) *satisfies*

$$\limsup_{t\to+\infty} x(t) \geq \xi^- \quad a.s. \tag{8}$$

and

$$\liminf_{t\to+\infty} x(t) \leq \xi^+ \quad a.s.\,, \tag{9}$$

where ξ^- *and* ξ^+ *are, respectively, the smaller and the larger positive roots of the function* $F : \mathbb{R}^+ \to \mathbb{R}$ *in* (7). *In other words,* $x(t)$ *will rise to or above* $\xi^- > 0$ *infinitely often with probability one.*

For a proof of the preceding result see [26, 27].

The next theorem states that solutions of (2) with a positive initial condition define a homogeneous Markov process. Its proof uses standard arguments in the stochastic differential equations literature (see, e.g. [22, 25]).

Theorem 6 *Let* $x(t) \in \mathscr{M}$ *be a solution with positive initial condition of the stochastic differential equation*

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + \sigma(x(t))^\alpha \mathrm{d}W_t, \quad t \geq 0$$

where σ *and* α *are positive constants and* f *satisfies the conditions of Corollary 1. Then* $x(t)$ *is a homogeneous Markov process, i.e. its transition probability*

$$P(y, s; A, t) = \mathbb{P}\left\{x_{y,s}(t) \in A\right\} \tag{10}$$

is such that

$$P(y, s; A, s + t) = P(y, 0; A, t)\,,$$

where $x_{y,s}(t)$ *is the solution of*

$$x_{y,s}(t) = y + \int_s^t f\left(x_{y,s}(u)\right) \mathrm{d}u + \int_s^t \sigma \left(x_{y,s}(u)\right)^{\alpha} \mathrm{d}W_u, \quad on\, t \geq s\ .$$

Let $\mathbb{P}_{x_0,t}(\cdot)$ denote the probability measure induced by a solution $x(t)$ of (2) with positive initial condition $x(0) = x_0$, that is

$$\mathbb{P}_{x_0,t}(A) = \mathbb{P}(x_{x_0,0}(t) \in A), \quad A \in \mathscr{B}(\mathbb{R}^+)\ .$$

If there is a probability measure $\mathbb{P}_\infty(\cdot)$ in $\left(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+)\right)$ such that

$$\mathbb{P}_{x_0,t}(\cdot) \to \mathbb{P}_\infty(\cdot) \quad \text{in distribution for any} \quad x_0 \in \mathbb{R}^+\ ,$$

we say that the SDE (2) has a stationary measure $\mathbb{P}_\infty(\cdot)$.

The next theorem gives conditions under which the SDE (2) has a unique stationary measure. For a proof see [26, 27].

Theorem 7 *Assume that one of the following sets of conditions holds:*

- $\alpha > 1$ *and* $n > 2$
- $\alpha > 1$, $n = 2$ *and* $\rho > \varepsilon$
- $\alpha = 1$, $n > 2$ *and* $\rho > \sigma^2/2$
- $\alpha = 1$, $n = 2$ *and* $\rho > \varepsilon + \sigma^2/2$.

Then, the SDE (2) *has a unique stationary distribution* $\mathbb{P}_\infty(\cdot)$ *in* $\left(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+)\right)$.

It should be noticed that the set of solutions for which there is a stationary measure is the same that guarantees persistence of the population.

2.5 Asymptotic Behaviour and the Stochastic Bifurcation Diagram

The evolution of the transition probability of our model is described by the Forward Kolmogorov equation associated with the SDE (2), namely

$$\frac{\partial}{\partial t} p(x,t) = -\frac{\partial}{\partial x}\left(f(x)p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(g^2(x)p(x,t)\right)\ , \tag{11}$$

where $f(x)$ and $g(x)$ are as given in (3) and (4), respectively. As we have already seen in the previous section, if $\mathbb{P}_\infty(\cdot)$ is a stationary measure in $\left(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+)\right)$ for the Markov process defined by the solutions of (2), then its density with respect to the Lebesgue measure in $\mathbb{R}^+$ is a steady state of (11).

Moreover, the unique steady state $p^{stat}(x)$ of (11) satisfying the constraints that $\mathrm{supp}(p^{stat}(x)) \subseteq \mathbb{R}^+$, $p^{stat}(x) \geq 0$ for every $x \in \mathbb{R}^+$ and

$$\int_{\mathbb{R}^+} p^{stat}(x)\mathrm{d}x = 1\ ,$$

is the density of the stationary measure $\mathbb{P}_\infty(\cdot)$, whenever such measure exists.

Computing the steady states of the Kolmogorov equation we obtain that, up to a multiplicative constant, these are given by

$$p_{\varepsilon,n}(x;\rho,\sigma,K,\alpha) = p_0(x;\rho,\sigma,K,\alpha)\mathrm{e}^{-\frac{2\varepsilon}{\sigma^2}E_{\alpha,n}(x)} ,$$

where

$$p_0(x;\rho,\sigma,K,\alpha) = x^{-2\alpha}\exp\left(\frac{-2\rho}{\sigma^2 K}\left(\frac{x^{3-2\alpha}}{3-2\alpha} - K\frac{x^{2-2\alpha}}{2-2\alpha}\right)\right) ,$$

for $\alpha \in (1,3/2)\cup(3/2,+\infty)$,

$$p_0(x;\rho,\sigma,K,1) = x^{-\left(2-2\rho/\sigma^2\right)}\exp\left(-2x\rho/\sigma^2 K\right) ,$$

$$p_0\left(x;\rho,\sigma,K,\frac{3}{2}\right) = x^{-\left(3+2\rho/\sigma^2 K\right)}\exp\left(-2\rho/\sigma^2 x\right)$$

and

$$E_{\alpha,n}(x) = \int_1^x \frac{y^{n-1-2\alpha}}{1+y^{n-1}}\mathrm{d}y , \quad x > 0 .$$

Although the expressions for $p_{\varepsilon,n}$ given above may look rather complicated, they are still amenable for analysis. Gathering the information provided by the explicit knowledge of the steady states of (11) and the previous results concerning extinction and persistence, we were able to construct a stochastic bifurcation diagram summarizing all the possible asymptotic behaviours of (2).

Let us start by discussing how the stationary measures change while the parameter $\alpha > 0$ changes (see Fig. 1). From Theorem 2, we know that whenever $\alpha < 1$, the solutions of (2) reach zero in finite time with full probability. Thus, there can be no absolutely continuous stationary measure when $\alpha < 1$. Instead, the Dirac measure based at zero is the unique possible stationary measure. If $\alpha = 1$, a very rich

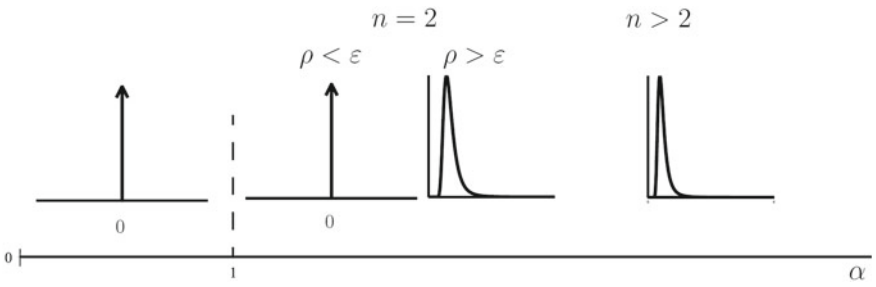


Fig. 1 Bifurcation Diagram for varying values of $\alpha > 0$

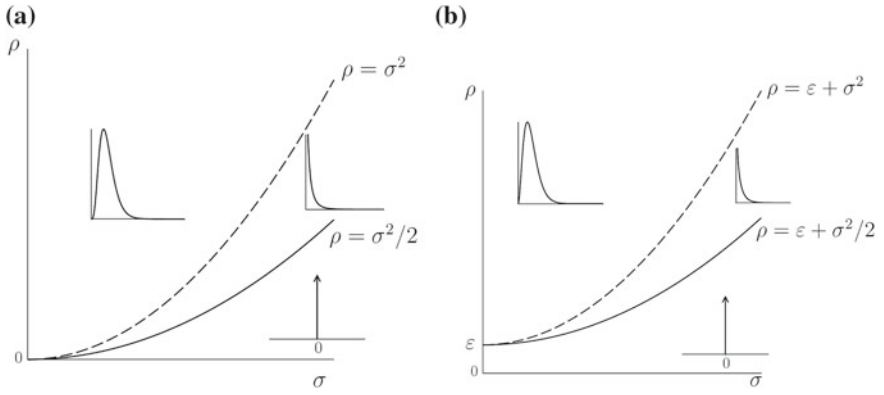


Fig. 2 **a** Bifurcation Diagram for varying values of ρ and σ in the case where $\alpha = 1$ and $n > 2$; **b** Bifurcation Diagram for varying values of ρ and σ in the case where $\alpha = 1$ and $n = 2$

picture containing several distinct qualitative behaviours emerges. We will discuss this particular case with detail in a moment.

When $\alpha > 1$, two distinct situations may occur, depending on whether $n = 2$ or $n > 2$. If $\alpha > 1$ and $n > 2$ there exists an absolutely continuous stationary measure whose density is a unimodal map. On the other hand, if $\alpha > 1$ and $n = 2$, then two distinct behaviours may occur. If $\rho < \varepsilon$, then by Theorem 3 the population becomes extinct with positive probability for a set of initial conditions with positive Lebesgue measure, i.e. an (eventual) stationary measure must contain a Dirac mass at zero. If however $\rho > \varepsilon$, then there exists an absolutely continuous stationary measure with an unimodal density.

Let us now consider the special case where $\alpha = 1$, i.e. the diffusion coefficient depends linearly on the population size (see Fig. 2). The cases $n = 2$ and $n > 2$ are again somewhat different. If $n > 2$ (resp. $n = 2$) and $\rho < \sigma^2/2$ (resp. $\rho < \varepsilon + \sigma^2/2$) then the population becomes extinct with positive probability for a set of initial conditions with positive Lebesgue measure and an (eventual) stationary measure must contain a Dirac mass at zero. At the bifurcation value $\rho = \sigma^2/2$ (resp. $\rho = \varepsilon + \sigma^2/2$), based on the divergence of the integral of $p_{\varepsilon,n}$ over $\mathbb{R}^+$, we conjecture that no absolute continuous stationary measure exists. However, for $\rho > \sigma^2/2$ (resp. $\rho > \varepsilon + \sigma^2/2$) an absolute continuous stationary measure exists. The stationary measure density is a strictly decreasing function if $\sigma^2/2 < \rho < \sigma^2$ (resp. $\sigma^2/2 < \rho - \varepsilon < \sigma^2$) and a unimodal function with limit zero when x tends to zero if $\rho > \sigma^2$ (resp. $\rho > \varepsilon + \sigma^2/2$). Finally, at the bifurcation value $\rho = \sigma^2$ (resp. $\rho = \varepsilon + \sigma^2$) the density $p_{\varepsilon,n}^{stat}$ is such that its limits as x tends to zero is finite and strictly positive.

3 Optimal Harvesting for a Stochastic Logistic Growth Model

We will now discuss the optimal harvesting policies associated with a population whose size evolves according to the stochastic logistic growth model described in Sect. 2 for the case where $\alpha \geq 1$ so that population extinction is not automatically guaranteed. Since the stochastic differential equation associated with this model does not always fit the standard assumptions in the stochastic optimal control literature, namely sublinear growth, we develop a dynamic programming principle for the stochastic optimal control problem we are interested in. We then use these results to provide a description of the optimal harvesting policies, as well as some qualitative properties of the corresponding value function.

3.1 Setup and Problem Formulation

Let $T > 0$ be some deterministic horizon. We will consider the random dynamical system defined by the following controlled stochastic differential equation

$$\begin{aligned} \mathrm{d}x(t) &= \left[\rho x(t)\left(1 - \frac{x(t)}{K}\right) - \varepsilon \frac{(x(t))^{n-1}}{1 + (x(t))^{n-1}} - h(t)x(t)\right]\mathrm{d}t + \sigma(x(t))^{\alpha}\mathrm{d}W_t \\ x(0) &= y\,, \quad y \geq 0\,, \end{aligned} \tag{12}$$

corresponding to the stochastic logistic growth model with predation of the previous section but with an extra term $h(t)x(t)$ representing the amount of harvesting to which the population is subjected to.

Our aim is to maximize the objective functional

$$J(y; h(\cdot)) = E\left[\int_0^T U(t, h(t))\mathrm{d}t + \Psi(T, x(T))\right] \tag{13}$$

subject to the stochastic dynamics determined by the SDE (12). The functions $U(t, \cdot)$ and $\Psi(T, \cdot)$ in the objective functional (13) are usually referred to as, respectively, the "profit rate" function and the "bequest" function. These functions are assumed to be strictly concave and increasing and represent, respectively, the utility derived from harvesting the population at a rate $h(t)$ throughout the interval $[0, T]$ and reaching the final time horizon T with a population of size $x(T)$. We will assume that the functions $U : [0, T] \times \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}$ and $\Psi : \mathbb{R}_0^+ \to \mathbb{R}$ are such that

- U and Ψ are uniformly continuous
- U and Ψ are Lipschitz continuous in x
- $U(t, x, \cdot)$ is a C^2 strictly increasing and strictly concave function for every $t \in [0, T]$ and $x \geq 0$
- $\Psi(\cdot)$ is a C^2 strictly increasing and strictly concave function.

In the next section we will use dynamic programming techniques to study the stochastic optimal harvesting problem (12) and (13). Afterwards, in Sect. 3.3, we will provide a detailed description of the optimal strategies associated with (12) and (13).

3.2 Dynamic Programming Principle and HJB Equation

The goal of this section is to obtain a dynamic programming principle for the stochastic optimal control problem defined by (12) and (13) and to derive the associated HJB equation.

Let $x_{0,y}(t; h(\cdot))$ denotes the state trajectory, starting from y when $t = 0$, associated with a control trajectory $h(\cdot)$. Denote by $\mathscr{A}^s[0, T]$ the set of *strong admissible control processes*, i.e. measurable and $\{\mathscr{F}_t\}$-adapted processes $h : [0, T] \times \Omega \to \mathbb{R}_0^+$ such that the stochastic differential equation (12) has a unique strong solution and the following integrability conditions hold:

$$\mathbb{E}\left[\int_0^T \left|U(t, x_{0,y}(t; h(\cdot)), h(t))\right| \mathrm{d}t\right] < \infty, \quad \mathbb{E}\left[\left|\Psi(x_{0,y}(T; h(\cdot)))\right|\right] < \infty .$$

The stochastic optimal control problem under consideration here amounts to find $h^*(\cdot) \in \mathscr{A}^s[0, T]$ maximizing the objective functional $J(y; h(\cdot))$ subject to the state equation (12) over the set of admissible controls $\mathscr{A}^s[0, T]$, that is

$$J(y; h^*(\cdot)) = \sup_{h(\cdot)\in\mathscr{A}^s[0,T]} J(y; h(\cdot)) . \tag{14}$$

The Markovian property from Theorem 6 makes the dynamic programming method particularly suitable to address this problem. Indeed, it enables a reduction of the initial optimal control problem to a two-parameter family of related problems, from which it is possible to extract a recursive relation leading to Bellman's optimality principle and the HJB equation. In order to proceed, we need to consider the weak formulation of the stochastic control problem, under consideration here as an auxiliary tool.

For any $(s, y) \in [0, T) \times \mathbb{R}_0^+$, consider the stochastic differential equation:

$$\begin{cases} \mathrm{d}x(t) = f(x(t), h(t))\mathrm{d}t + \sigma(x(t))^{\alpha}\mathrm{d}W_t , \quad t \in [s, T] \\ x(s) \quad = y \end{cases} \tag{15}$$

together with the objective functional

$$J(s, y; h(\cdot)) = \mathbb{E}\left[\int_s^T U(t, x_{s,y}(t; h(\cdot)), h(t))\mathrm{d} + \Psi\left(x_{s,y}(T; h(\cdot))\right)\right] , \tag{16}$$

where f and g are as given in (3) and (4), and $x_{s,y}(t; h(\cdot))$ is the solution of (15) associated with the control $h(\cdot)$ and starting from y when $t = s$.

Let $\mathscr{A}^w[s, T]$ denote the set of weak admissible controls and note that for any $(s, y) \in [0, T) \times \mathbb{R}_0^+$ and $h(\cdot) \in \mathscr{A}^w[s, T]$, the SDE (15) admits a unique solution $x(\cdot) = x_{s,y}(\cdot; h(\cdot))$. Hence, the objective functional (16) is well-defined. Moreover, the *value function* $V : [0, T] \times \mathbb{R}_0^+ \to \mathbb{R}$ is well-defined as

$$\begin{cases} V(s, y) = \sup_{h(\cdot)\in\mathscr{A}^w[s,y]} J(s, y; h(\cdot)) \\ V(T, y) = \Psi(y) \end{cases} . \tag{17}$$

The following two results provide implicit descriptions for the value function V defined above: the dynamic programming principle and the corresponding HJB equation. For further details and their proofs see [26, 29].

Theorem 8 (Bellman's Optimality Principle) *For any* $(s, y) \in [0, T) \times \mathbb{R}_0^+$ *and* $s \leq s' \leq T$ *we have that*

$$V(s, y) = \sup_{h(\cdot)\in\mathscr{A}^w[s,T]} \mathbb{E}\Bigg[V(s', x_{s,y}(s'; h(\cdot))) + \int_s^{s'} U(t, x_{s,y}(t; h(\cdot)), h(t))\mathrm{d}t \Bigg] .$$

This theorem gives a backwards recursive relation for the function V that can be used to obtain a HJB equation, a partial differential equation whose solution, whenever exists, is the value function of the optimal control problem under consideration here. Let $I \subseteq \mathbb{R}$ be an interval and denote by $C^{1,2}\left(I \times \mathbb{R}_0^+; \mathbb{R}\right)$ the set of all continuous functions $V : I \times \mathbb{R}_0^+ \to \mathbb{R}$ such that V_t, V_x, and V_{xx} are all continuous functions of $(t, x) \in \overset{\circ}{I} \times \mathbb{R}^+$.

Theorem 9 (Hamilton Jacobi–Bellman equation) *If the value function* V *is such that* $V \in C^{1,2}\left([0, T) \times \mathbb{R}_0^+; \mathbb{R}\right)$, *then it satisfies the boundary value problem*

$$\begin{cases} V_t + \sup_{h\in\mathbb{R}_0^+} \mathscr{H}(t, x, h, V_x, V_{xx}) = 0 \\ V(T, x) = \Psi(T, x) \end{cases} \tag{18}$$

where the Hamiltonian function is given by

$$\mathscr{H}(t, x, h, V_x, V_{xx}) = U(t, x, h) + f(x, h)V_x + \frac{1}{2}\sigma^2 x^{2\alpha} V_{xx} , \tag{19}$$

and f *and* g *are given in* (3) *and* (4), *respectively.*

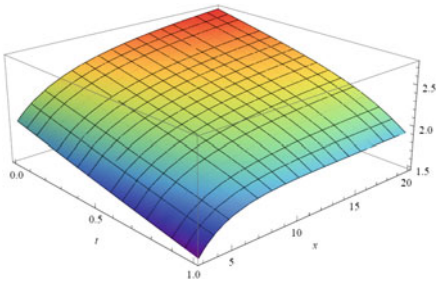


Fig. 3 Value Function V for the set of parameters $(\mathbf{P}) : T = 1,\ \rho = 1,\ K = 10,\ \varepsilon = 0.5,\ n = 3,\ \sigma = 0.1,\ \alpha = 1.5,\ \gamma = 0.5,\ \beta = 0.5,\ \theta = 0.04$

3.3 The Optimal Harvesting Problem

In this section we will provide some qualitative properties of the value function defined by (17). We also provide a characterization of the optimal strategy for the particular case of Constant Absolute Risk Aversion Utilities.

Proposition 2 *The value function V is strictly increasing with respect to x. Furthermore, V is increasing with respect to the carrying capacity K and decreasing with the predation size ε.*

The proof of the previous result can be found in [26, 29]. For a plot of the value function for a specific choice of model parameters see Fig. 3.

3.3.1 The Case of Constant Absolute Risk Aversion Utilities

We will now further specialize our discussion to the class of discounted exponential utility functions

$$U(t,h) = \mathrm{e}^{-\theta t}\frac{1-\mathrm{e}^{-\gamma h}}{\gamma}, \quad \Psi(T,x) = \mathrm{e}^{-\theta T}\frac{1-\mathrm{e}^{-\beta x}}{\beta}, \tag{20}$$

where the risk aversion parameters γ and β, as well as the discount rate θ, are strictly positive constants.

The family of utility functions in (20) has the property of having a constant Arrow-Pratt coefficient of absolute risk aversion (firstly introduced in [2, 30]), making these utility functions key examples for the modelling of preference relations in Economic Theory [23].

Combining the maximizer of the Hamiltonian function H given in (19) with (20), we obtain that the optimal harvesting strategy is given by

$$h^*(t,x) = -\frac{1}{\gamma}\left(\theta t + \ln(x V_x)\right)\ . \tag{21}$$

Substituting the optimal harvesting $h^*(t,x)$ above in the HJB equation, we arrive at the nonlinear second order partial differential equation

$$V_t + \frac{1}{\gamma}\left(\mathrm{e}^{-\theta t} - (1 - \theta t - \ln(xV_x))xV_x\right) + \left(\rho x\left(1 - \frac{x}{K}\right) - \varepsilon\frac{x^{n-1}}{1+x^{n-1}}\right)V_x + \frac{1}{2}\sigma^2 x^{2\alpha}V_{xx} = 0 \tag{22}$$

with terminal condition given by

$$V(T,x) = \mathrm{e}^{-\theta T}\frac{1-\mathrm{e}^{-\beta x}}{\beta}\,. \tag{23}$$

We will now list the conclusions of a static analysis for the optimal harvesting strategies associated with the optimal control problem under consideration here. Such analysis is based on small variations on the set of parameters used for the construction of Fig. 3.

We start by noticing that the optimal harvesting is increasing as a function of both t and x, being also convex with respect to time. See Fig. 4.

In what concerns the remaining model parameters, experiments performed by numerical integration of the PDE (22) and (23) indicate that h^* increases with:

(i) increasing values of natural growth rate ρ
(ii) decreasing values of the predation size ε and decreasing values of the Holling functional parameter n
(iii) increasing values of the volatility coefficient σ and of the convexity parameter α
(iv) decreasing values of the risk aversion parameters γ and β and of the discount rate θ.

In what concerns the parameters σ, α and n, we should also add that these seem to have very little influence on the feedback form of the optimal strategies.

Finally, we remark that the numerical results listed above seem to be robust with respect to realistic changes in the parameters.

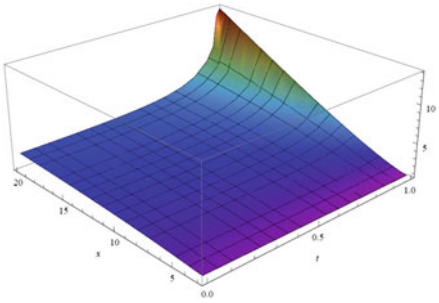


Fig. 4 Optimal harvesting h^* in feedback form for the set of parameter values **(P)**

4 Conclusion

We have provided a detailed description of the dynamics of a stochastic logistic growth model with a predation term and a diffusion coefficient of power type. The nonlinearity introduced by the diffusion coefficient leads to interesting distinct asymptotic behaviours depending on the convexity of such coefficient. Another key ingredient of the stochastic logistic growth model under consideration here is the presence of a predation term given by a Holling type-n functional response. The combined influence of the population natural growth rate and the size of the predation term turn out to be responsible for some of the different qualitative behaviours described here.

We have also studied an optimal harvesting problem associated with this stochastic logistic growth model in the case where $\alpha \geq 1$. As a preliminary step, and since our SDE model does not always fit the standard assumptions in the stochastic optimal control literature (i.e. sublinear growth), we have provided a dynamic programming principle for the stochastic optimal control problem we are interested in. We then use such results to proceed with a static analysis of the optimal harvesting strategies in the case where the utilities belong to the family of constant absolute risk aversion utilities.

Acknowledgements The author thanks Diogo Pinheiro, Alberto Pinto and Athanasios Yannacopoulos for their comments and advice. Part of the research leading to this paper was done during visits to Athens University of Economics and Business, whom the author thanks for the hospitality. S. Pinheiro's research was supported by FCT - Fundação para a Ciência e Tecnologia grant with reference SFRH/BD/61547/2009.

References

1. Alvarez, L.H.R., Shepp, L.A.: Optimal harvesting of stochastically fluctuating populations. J. Math. Biol. **37**(2), 155–177 (1998)
2. Arrow, K.J.: In: Säätiö, Y.J. (ed.) The Theory of Risk Aversion. In Aspects of the Theory of Risk Bearing. Helsinki (1965)
3. Beddington, J.R., May, R.M.: Harvesting natural populations in a randomly fluctuating environment. Science **197**, 463–465 (1977)
4. Bellman, R.E.: On the theory of dynamic programming. Proc. Nat. Acad. Sci. USA **38**, 716–719 (1952)
5. Bellman, R.E.: An introduction to the theory of dynamic programming. Rand Corporation Report, R–245 (1953)
6. Bellman, R.E.: Dynamic programming and a new formalism in the calculus of variations. Proc. Nat. Acad. Sci. USA **40**, 231–235 (1954)
7. Bellman, R.E.: Dynamic programming and stochastic control process. Inf. Control **1**, 228–239 (1958)
8. Braumann, C.A.: Variable effort fishing models in random environments. Math. Biosci. **156**, 1–19 (1999)
9. Braumann, C.A.: Variable effort harvesting models in random environments: generalization to density-dependent noise intensities. Math. Biosci. **177–178**, 229–245 (2002)
10. Clark, C.: Mathematical Bioeconomics. Wiley-Interscience Publication (1931)

11. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions, 2nd edn. Springer, New York (2006)
12. Florentin, J.J.: Optimal control of continuous-time Markov stochastic systems. J. Electron. Control **10**, 473–488 (1961)
13. Florentin, J.J.: Partial observability and optimal control. J. Electron. Control **13**, 263–279 (1962)
14. Gleit, A.: Optimal harvesting in continuous time with stochastic growth. Math. Biosci. **41**, 112–123 (1978)
15. Gray, A., Greenhalgh, D., Hu, L., Mao, X., Pan, J.: A stochastic differential equation SIS epidemic model. SIAM J. Appl. Math. **71**(3), 876–902 (2011)
16. Ji, C., Jiang, D., Shi, N., O'Regan, D.: Existence, uniqueness, stochastic persistence and global stability of positive solutions of the logistic equation with random perturbation. Math. Methods Appl. Sci. **30**, 77–89 (2007)
17. Jiang, D., Shi, N.: A note on nonautonomous logistic equation with random perturbation. J. Math. Anal. Appl. **303**, 164–172 (2005)
18. Jiang, D., Shi, N., Li, X.: Global stability and stochastic permanence of a non-autonomous logistic equation with random perturbation. J. Math. Anal. Appl. **340**, 588–597 (2008)
19. Kushner, H.J.: Optimal stochastic control. IRE Trans. Auto. Control **AC–7**, 120–122 (1962)
20. Lande, R., Engen, S., Saether, B.E.: Optimal harvesting of fluctuating populations with a risk of extinction. Am. Nat. **145**(5), 728–745 (1995)
21. Lungu, E.M., Øksendal, B.: Optimal harvesting from a population in a stochastic crowded environment. Math. Biosci. **145**(1), 47–75 (1997)
22. Mao, X.: Stochastic Differential Equations and Application. Horwood Publishing (2007)
23. Mas-Collel, A., Whinston, M.D., Green, J.: Microeconomic Theory. Oxford University Press, Oxford (1995)
24. May, R.M., Beddington, J.R., Horwood, J.H., Shepherd, J.G.: Exploiting natural populations in an uncertain world. Math. Biosci. **42**, 219–252 (1978)
25. Øksendal, B.: Stochastic Differential Equations: An Introduction with Application. Springer, Berlin (2010)
26. Pinheiro, S.: Dynamics of holomorphic and stochastic differential equations. Ph.D. thesis, Universidade do Porto (2014)
27. Pinheiro, S.: On a logistic growth model with predation and a power-type stochastic perturbation I (Submitted)
28. Pinheiro, S.: On a logistic growth model with predation and a power-type stochastic perturbation II (Submitted)
29. Pinheiro, S.: Optimal harvesting for a logistic growth model with predation and a power-type diffusion coefficient (Submitted)
30. Pratt, J.W.: Risk aversion in the small and in the large. Econometrica **32**, 122–136 (1964)
31. Roberts, M.G., Saha, A.K.: The asymptotic behaviour of a logistic epidemic model with stochastic disease transmission. Appl. Math. Lett. **12**, 37–41 (1999)
32. Yong, J., Zhou, X.Y.: Stochastic Controls Hamiltonian Systems and Equations. Springer, Berlin (1999)

Myopia of Governments and Optimality of Irreversible Pollution Accumulation

Laura Policardo

Abstract In this paper I address the question of whether irreversible pollution accumulation - in a global pollution problem - may be optimal or not. Based on the Tahvonen and Withagen's article (Tahvonen, Withagen, J Econ Dyn Control, 20:1775–1795, 1996), [16], I set up a model of economic growth where pollution is a byproduct of production, and its natural decay function follows an inverted-U shape, and becomes irreversible for high levels of pollution. Under some parameter's constellation, the model produces multiplicity of equilibria making local analysis of little relevance. I therefore study the global dynamics of the system using a dynamic programming algorithm, showing that irreversible pollution accumulation cannot be an optimal strategy, unless it is guided by short-term objectives.

Keywords Economic growth · Irreversible pollution accumulation · Dynamic programming · Global dynamics · Multiplicity of equilibria

JEL Classification E27 · E61 · O13 · O21 · Q58

1 Introduction

Despite several effects of pollution - like global warming - have a global nature, involving all the countries irrespective of who is responsible for producing wastes, environmental policies are decided in autonomy by the single nations. The coordination problem underlying this "hot" topic is one of the main reasons of the steady growth of greenhouse gases and other toxic substances, which the scientific consensus believes they are the main causes leading to global warming.

L. Policardo (✉)
University of Siena, Siena, Italy
e-mail: policardo@unisi.it

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_17

Hoping to make a contribution towards the awareness of the necessity of having a unique, global, environmental policy, in this paper I study a model of economic growth with pollution accumulation, where pollution has a nonlinear decay function which follows an inverted-U shape and becomes irreversible when a given stock is reached. In the analysis, I will focus on the case of an efficiently and infinitely lived planned economy, with the aim at responding to the question of whether irreversible pollution accumulation may be an optimal strategy or not.

The model I use in the paper is built on the basis of Tahvonen and Withagen's one, published on the JEDC in 1996 [16]. I generalise their model by introducing a capital accumulation function and assuming a global pollution problem instead of a local one. Globality of the problem is reflected in the introduction of the hypothesis of a "subsistence" level of consumption. This assumption is crucial in determining the dimensionality of the problem, since in my model the population cannot leave to move in a cleaner and unpolluted area.

Since this problem produces multiplicity of stable solutions, local analysis gives little insights since it does not allow to say what is the dynamics between different equilibria, or far away from them. In order to fill this informational gap, I study the global dynamics of the model using a dynamic programming algorithm (carefully explained in the appendix, with codes also included), and comparing different paths in terms of welfare they produce.

In the literature of economic growth and the environment, little attention has been paid to the fact that the natural decay function of pollution might be endogenously determined by the stock of pollution itself. Linearity has been a commonly assumed hypothesis, and that allowed economists to find an unique stable stationary solution of the system (Keeler E., Spence M., and Zeckhauser R. [9], Nancy Stokey [15]). The main consequence of this hypothesis was that the insights one could learn from these lessons were that in the long run, all the countries would have converged to that unique equilibria, without realising that the hypothesis they implicitly made was that the more polluted were the environment, the more it was able to clean itself up, hypothesis quite unrealistic. Moreover, this prediction is clearly in sharp contrast with the evidence today, where many countries still find the joining to international protocols not worthwhile, and keep their emissions' level unbounded. The prediction of uniqueness of equilibria is a direct consequence of the choice of the decay function of pollution, because using different function of pollution's decay may lead to a different solution specifications, even to a multiplicity of stationary solutions.

The choice of such inverted-U shaped decay function for pollution is due mainly to the observation of natural phenomena, which suggest - contrary to what is commonly assumed in the economic literature - that the natural self recreation capacity of the environment certainly isn't always increasing with respect to the stock of pollution. A sort of endogeneity of this ability of the environment was first noticed by Holling [8], who, in an article published in 1972, wrote about the fact that nutrient enrichment of lakes changed its biodiversity permanently, making the lake incapable of recovery its original status even if emissions were to be eliminated. Several authors, subsequently,

raised the problem (Dasgupta [4], Forster [7]). Recently, in his highly debated report, Stern [14] predicted future scenarios where, if emissions are kept at the "business as usual" level, global warming may change dramatically the biodiversity of the planet through desertification and the rise of sea levels, putting at serious risk people's health and lowering the probability of survival for some other populations, and therefore suggesting a sort of inability of the environment to absorb pollution, for high pollution levels. Despite all these contributions, this fact has gained very little relevance in the economic literature.

The paper is organised as follows: Sect. 2 introduces the theoretical model and describes its properties, Sect. 3 introduces the strategies to study the local and global dynamics of this system and presents the results, and Sect. 4 concludes.

2 The Model

In this section, I introduce the model of economic growth where citizens' utility depends on both consumption and pollution, and technology is linear. The function for pollution accumulation depends on both the level of production, and its natural rate of decay, which is endogenous to the stock of pollution and follows an inverted-U shape. For simplicity, I describe an optimal solution dictated by a benevolent planner who acts in the interest of the citizens (in the following, I will use interchangeably the words "citizen", "household" and "representative agent" since all these interpretations are correct and do not change the scope of this study). The citizens' instantaneous utility is a separable function of consumption,[1] denoted by C, and pollution,[2] x, according to the following rule:

$$u(c, x) = v(C) - h(x) \tag{1}$$

with v strictly increasing and strictly concave, and h strictly increasing and strictly convex. C is composed by two arguments,

$$C(t) = c(t) + mc \tag{2}$$

with c the level of consumption beyond the minimum subsistence level denoted by mc, with $mc > 0$, that is kept constant at all the times. It follows that C is always positive and lower bounded by mc.

[1] This is a special assumption since this formulation implies that the enjoyment of consumption does not depend on pollution, and that the disutility of pollution is not affected by the level of consumption.

[2] Pollution in this model is a public "bad", so each individual experience its whole amount, while consumption is considered in percapita terms.

These four properties also hold:

$$\lim_{c\to 0} v'(c) = \overline{MU}_{mc} < \infty \tag{3}$$

$$\lim_{c\to \infty} v'(c) = 0 \tag{4}$$

$$\lim_{x\to 0} h'(x) = 0 \tag{5}$$

$$\lim_{x\to \infty} h'(x) = +\infty \tag{6}$$

where Eq. 3 represents the maximum achievable level of marginal utility from consumption, that is to say, the level of marginal utility at the subsistence or minimum level. Condition (4) indicates that the marginal utility from consumption decreases as consumption increases, converging to zero for levels of consumption tending to infinity, and conditions (5) and (6) indicate that at low levels of pollution, the marginal disutility is low, but increases as the stock of pollution increases. The representative agent discounts at the subjective discount rate all the future flows of utility, so his total welfare is

$$U(C, x) = \int_0^{\infty} u(C, x)e^{-\rho t}dt \tag{7}$$

Production is linear in the argument of capital, so capital's productivity is constant and equal to A:

$$y(t) = Ak(t) \tag{8}$$

Capital is then a necessary factor of production, and in order to guarantee a minimum subsistence level of consumption, it must be strictly positive, so as production. Assume that $\underline{k}$ is the minimum level of capital which guarantees a production equal to $\underline{y}$. The minimum level of production $\underline{y}$ must be such that the amount of investments is equal to the capital's depreciation, and the amount of consumption in each period is equal to mc. It follows that

$$k \in [\underline{k}, \infty), \qquad \underline{k} > 0 \tag{9}$$

$$y \in [\underline{y}, \infty), \qquad \underline{y} > 0 \tag{10}$$

The planner decides on behalf of the citizens how much production to consume and to invest to accumulate further capital. The capital accumulation function of the economy is represented by

$$\dot{k}(t) = y(t) - \delta k(t) - C(t) \tag{11}$$

with δ representing the constant rate of capital depreciation. Pollution is a byproduct of production, and it is assumed to obey the following equation

$$\dot{x}(t) = y(t) - \eta\left(x(t) - \frac{\theta}{\eta}x(t)^2\right) \quad \text{if } x(t) < \eta/\theta \tag{12}$$

$$\dot{x}(t) = y(t) \quad \text{if } x(t) \geq \eta/\theta \tag{13}$$

Equations 12 and 13 represent the law of pollution accumulation. The term $\eta\left(x(t) - \frac{\theta}{\eta}x(t)^2\right)$ is the pollution's natural decay, and follows, as anticipated in the introduction, an inverted-U shape. $\bar{x} = \eta/\theta$ represents the threshold beyond which pollution becomes irreversible (so the decay is zero).

In order to write down the conditions for maximization, I will use the same utility function used by Stokey [15] so the specification of the welfare function becomes:

$$v(C) = \frac{C(t)^{1-\sigma} - 1}{1 - \sigma} \tag{14}$$

$$h(x) = \frac{Bx(t)^{\gamma}}{\gamma} \tag{15}$$

with $\sigma > 0$, $B > 0$ and $\gamma > 1$. Moreover, I will assume in the following:

HP 2.1. The marginal product of capital net of the depreciation is positive, $A - \delta > 0$.

HP 2.2. The marginal product of capital, net of the depreciation is greater that the intertemporal rate of preferences, $A - \delta > \rho$

HP 2.3. The sum of the intertemporal rate of preferences and the marginal rate of decay of pollution is positive, $\rho + \eta(1 - 2 \cdot \frac{\theta}{\eta}x) > 0$. As long as the marginal decay function is positive, this hypothesis is always satisfied, but when it is negative, this implies that the rate of impatience is greater than the marginal loss in the self purification capacity of the environment.

Later, I will compare two possible outcomes of this model: a reversible solution and an irreversible one. In the first case, I will study the reversible solution, assuming that the planner will maximise utility in infinite time letting pollution to stay below its threshold level forever. In the second case, I will study an irreversible solution, and since the solution admits a point of non-differentiability, I will follow the same approach used by Tahvonen and Withagen and I will split the problem into two subproblems: a first period problem, where the planner maximises utility from zero to T (finite time) letting pollution to reach the irreversibility threshold at T, followed by a second period problem where the planner maximises utility from T to infinity when the natural decay function for pollution is nil.

2.1 Reversible Pollution Accumulation

Let us assume that the planner wants to maximise the representative citizen's welfare having an infinite time horizon plan, and letting pollution not to reach $\bar{x}$. In this case, the problem faced by the planner is:

$$\max_{c(t)} W^{\infty} = \int_0^{\infty} e^{-\rho t}\left[\frac{C(t)^{1-\sigma}-1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right]dt \tag{16}$$

subject to

$$\dot{k}(t) = (A-\delta)k(t) - C(t) \tag{17}$$

$$\dot{x}(t) = Ak(t) - \eta\big(x(t) + \frac{\theta}{\eta}x(t)^2\big) \tag{18}$$

$$\lim_{t\to\infty} x(t) < \bar{x} \tag{19}$$

Denote the solution of this first period problem by $(c^{\infty}, k^{\infty}, x^{\infty})$ and the respective costate variables by λ_1^{∞} and λ_2^{∞}. Denote also the flow of utility yield by this optimal plan W^{∞}. The Hamiltonian for this problem is

$$\mathscr{H}(t, k(t), x(t), c(t), \Lambda; \Theta) \overset{def}{=} \lambda_0 \cdot \left[\frac{C(t)^{1-\sigma}-1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right] + \\ +\lambda_1(t)\Big[(A-\delta)k(t) - c(t)\Big] + \lambda_2(t)\Big[Ak(t) - \eta\big(x(t) + \frac{\theta}{\eta}x(t)^2\big)\Big] \tag{20}$$

where Λ is the set of shadow prices, $\Lambda = \{\lambda_0(t), \lambda_1(t), \lambda_2(t)\}$ and Θ represents the set of exogenous parameters of the model, $\Theta = \{A, B, \sigma, \rho, \delta, \eta, \theta, \gamma, mc\}$ and, in more detail, $\lambda_1(t)$ represents the shadow price of capital, and $\lambda_2(t)$ the shadow price of pollution.

The maximum principle asserts that there exists a λ_0 and a continuous and piecewise continuously differentiable functions $\lambda_1(t)$ and $\lambda_2(t)$, such that for all t

$$(\lambda_0, \lambda_1(t), \lambda_2(t)) \neq (0,0,0) \tag{21}$$

$$\mathscr{H}(t, k^*(t), x^*(t), c^*(t), \Lambda; \Theta) \geq \mathscr{H}(t, k^*(t), x^*(t), c(t), \Lambda; \Theta) \qquad \forall t \tag{22}$$

The necessary first order conditions are:

$$\frac{\partial \mathscr{H}}{\partial c} = 0 \quad \Rightarrow \quad \lambda_1 = C^{-\sigma} \tag{23}$$

$$\frac{\partial \mathscr{H}}{\partial k} = \rho\lambda_1 - \dot{\lambda}_1 \quad \Rightarrow \quad \dot{\lambda}_1 = \lambda_1(\rho + \delta - A) - \lambda_2 A \tag{24}$$

$$\frac{\partial \mathscr{H}}{\partial x} = \rho\lambda_2 - \dot{\lambda}_2 \quad \Rightarrow \quad \dot{\lambda}_2 = \lambda_2 \cdot [\rho + \eta\big(1 - \frac{2\theta}{\eta}x\big)] + Bx^{\gamma-1} \tag{25}$$

$$\lambda_0 = 1 \quad \text{or} \quad \lambda_0 = 0 \tag{26}$$

and sufficient conditions for maximisation are the following transversality conditions:

$$\lim_{t\to\infty} e^{-\rho t}\lambda_1(t) \cdot k(t) = 0 \tag{27}$$

$$\lim_{t\to\infty} e^{-\rho t}\lambda_2(t)\cdot x(t) = 0 \tag{28}$$

$$\lambda_1(t) \geq 0 \tag{29}$$

$$\lambda_2(t) \leq 0 \tag{30}$$

Since the terminal conditions for capital and pollution as time approaches infinity are left free, it follows that $\lim_{t\to\infty}\lambda_1(t) = 0$ and $\lim_{t\to\infty}\lambda_2(t) = 0$ so necessarily, because of condition (26), $\lambda_0 = 1$.

The following system of four differential equations represents the conditions any optimal path has to obey:

$$\dot{k} = (A-\delta)k - \lambda_1^{-\frac{1}{\sigma}} \tag{31}$$

$$\dot{x} = Ak - \eta x + \theta x^2 \tag{32}$$

$$\dot{\lambda}_1 = \lambda_1(\rho - (A-\delta)) - \lambda_2 A \tag{33}$$

$$\dot{\lambda}_2 = \lambda_2(\eta - 2\theta x + \rho) + Bx^{\gamma-1} \tag{34}$$

with Eqs. 33 and 34 representing the Euler equations. In equilibrium, all the variables in the economy grow at a zero rate, so $\dot{k} = \dot{x} = \dot{\psi} = \dot{\lambda} = 0$.

I first start by analyzing the so-called corner solutions, that is to say solutions that assume consumption equal to the minimum subsistence level. Assuming

HP 2.4. $C^*(t) = mc$

and also

HP 2.5. $mc < \frac{\eta^2(A-\delta)}{4\theta A}$ (This condition is necessary to guarantee the level of pollution be real)

it follows that there are two simultaneous steady states represented in the table below:

Equilibrium 1	Equilibrium 2
$k^* = mc/(A-\delta)$	$k^* = mc/(A-\delta)$
$x^* = \frac{\eta-\sqrt{\eta^2-\frac{mc\cdot 4\theta A}{A-\delta}}}{2\theta}$	$x^* = \frac{\eta+\sqrt{\eta^2-\frac{mc\cdot 4\theta A}{A-\delta}}}{2\theta}$
$\lambda_1^* = mc^{-\sigma}$	$\lambda_1^* = mc^{-\sigma}$
$\lambda_2^* = -\frac{Bx_1^{*\gamma-1}}{\eta-2\theta x_1^*+\rho}$	$\lambda_2^* = -\frac{Bx_1^{*\gamma-1}}{\eta-2\theta x_1^*+\rho}$

Due to the inverse U-shaped function for the pollution decay, this corner solution admits two stationary points, for each value of mc respecting condition 2.5.

Now, I consider interior solutions. From Eq. 32, it is possible to see that considering $\dot{x} = 0$ and rearranging I get

$$k = \frac{x(\eta - \theta x)}{A} \tag{35}$$

and, combining equations 31, 33, 34 and considering $\dot{k} = \dot{\psi} = \dot{\lambda} = 0$ I get

$$k = \frac{1}{(A-\delta)} \cdot \left(\frac{AB}{A-\delta-\rho}\right)^{-\frac{1}{\sigma}} \cdot x^{\frac{1-\gamma}{\sigma}} \cdot (\eta - 2\theta x + \rho)^{\frac{1}{\sigma}} \tag{36}$$

Any intersection between the two Eqs. 35 and 36 represents an equilibria.[3] In general, the existence of an equilibria (or multiplicity of equilibria) depends on the choice of the parameters of the model. Equation 36 is decreasing in all his domain, whilst Eq. 35 has an inverted-U shape. Graphically, one may have the following cases:

The first graph represents a case where two interior solutions exist, and those are represented by $E1$ and $E2$. At the same time, this picture shows that there might exist two additional corner solutions, represented by the intersection between the horizontal line (which identifies the minimum level of capital that is necessary to guarantee a consumption equal to the subsistence level and to cover capital's depreciation). Those solutions are, respectively, $E3$ and $E4$.

The second picture depicts instead another case where there are still two interior solutions, but one (represented by $E2$) cannot be considered a feasible equilibria since its level of consumption is lower than the subsistence level. This case therefore leads to only three feasible stable solutions.

Case three represents a different situation where only one interior solution exists, with associated level of consumption higher than mc. Despite the fact that this solution requires a different parameter's set with respect the previous case, the outcome is similar since it generates three feasible steady solutions.

Case four happens when mc is larger than the equilibrium levels of all the interior solutions, but the two corner solutions still respect Proposition 2.3. Hence, the number of feasible stable solutions is only two and those are the corner solutions.

Case five occurs when mc is equal to $\frac{\eta^2(A-\delta)}{4\theta A}$. This situation leads to just one stable solution, irrespective of the number of the existing interior solutions. This is because if they existed, they would have necessary a level of equilibrium consumption necessarily lower than the subsistence level. Case six shows, instead, that whatever the number of interior solutions, if $mc > \frac{\eta^2(A-\delta)}{4\theta A}$, no feasible steady state can exist, because of the reason above.

It follows that the next propositions hold:

Proposition 2.1 *Necessary and sufficient condition to have one interior stable solution is* $\eta < \rho$.

Proposition 2.2 *Necessary and sufficient condition to have either two or zero interior stable solutions is* $\rho > \eta$, *sufficient condition to have two interior solutions is* $\rho < \Psi \cdot (\eta/\theta)^{2\sigma+\gamma-1}$ *with* $\lambda_1 = (1/2)^{\sigma+\gamma-1} \cdot (A-\delta)^{\sigma} \cdot (\frac{AB}{A-\delta-\rho})$

Proposition 2.3 *Necessary and sufficient condition to have two (one) corner solution(s) is* $mc < (=)\eta^2(A-\delta)/(4A\theta)$.

[3] This rearrangement of Eqs. 31–34 is only aimed at expressing the two stationary solutions in the $k - x$ plane and Eqs. 35 and 36 do not have necessarily an economic interpretation.

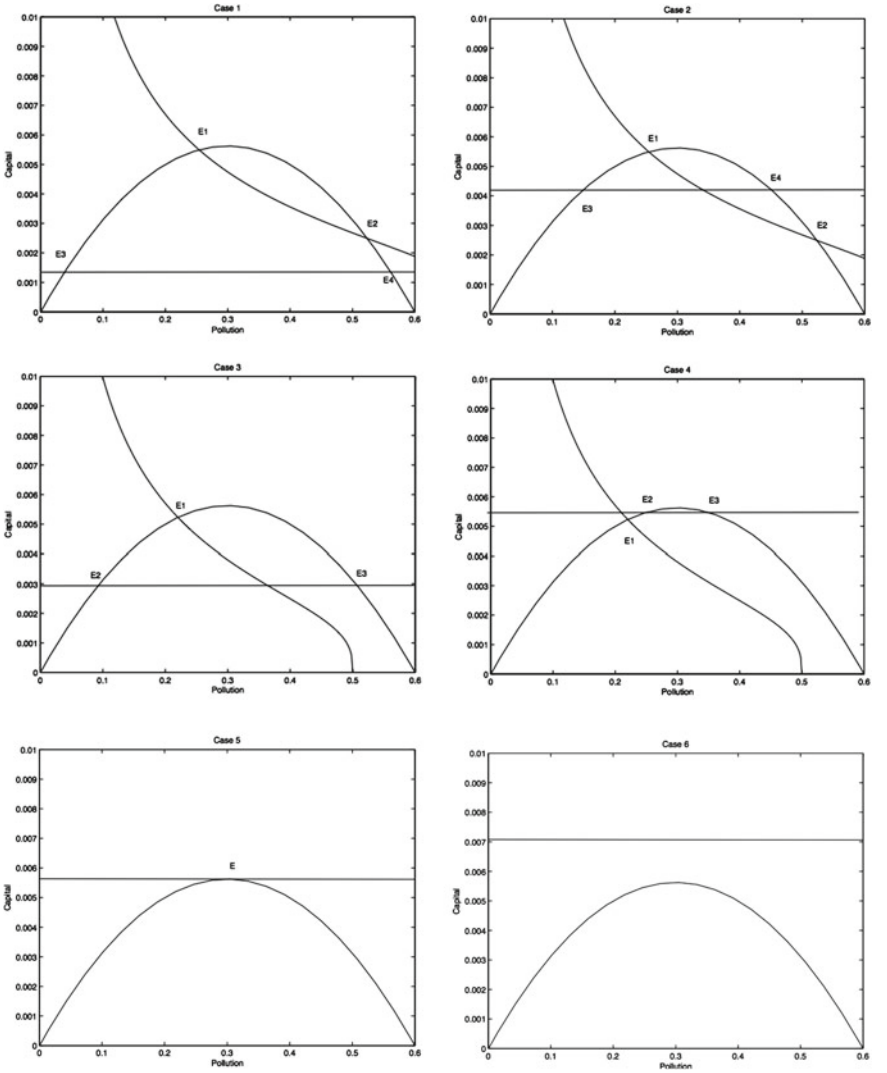


Fig. 1 *Graphical representation of different cases, ranging from four to zero steady states. These cases are not exhaustive.*

The first two sets of graphs (case 1 and 2) have been obtained using the following set of parameters: $B = 10,000,000$, $A = 0.8$, $\rho = 0.04$, $\theta = 0.05$, $\eta = 0.03$, $\sigma = 3$, $\gamma = 3$, and $\delta = 0.1$. The second two graphs (case 3 and 4) have instead been obtained using $B = 10,000,000$, $A = 0.8$, $\rho = 0.02$, $\theta = 0.05$, $\eta = 0.03$, $\sigma = 3$, $\gamma = 3$, and $\delta = 0.1$. Irrespective of the set of parameters chosen, the last two graphs suggest that if the level of capital at the minimum level of consumption is as high as the maximum level of capital corresponding to the turning point of the decay function for pollution, we can only have one steady state, and if it is higher, no steady states at all. The last two graphs, instead, do not respect H.2.5 because, in graph 5, $mc = \frac{\eta^2(A-\delta)}{4\theta A}$ and in graph 6 $mc > \frac{\eta^2(A-\delta)}{4\theta A}$, with the consequence, respectively, of the existence of only one (or two equal) solutions for pollution, or zero

Depending on the choice of the parameters, I can have up to a maximum of four different equilibria. Consider for example case 1 in Fig. 1, and assume that $\underline{k}$ is lower than the level of capital in equilibrium E2. This implies the existence of four non-trivial stationary solutions. On the other hand, however, if the level of $\underline{k}$ is higher than the level of capital in equilibrium at E2, E2 cannot be considered a valid solution and therefore the number of equilibria are three.

In what follows, I will confine my analysis to the case where multiplicity of steady states occurs because, from an economic point of view, I believe it is the most interesting and the most realistic. It is not unusual indeed to see different countries with characteristics that can be represented by such a configuration of stable solutions. For instance, it is generally agreed that the cleanest cities in the world are located in developed and rich countries, like Canada, Finland, Norway etc. The worst polluted countries are mainly in China and India, that although they are growing at very high rate, they are not certainly rich countries. On the converse, there are natural paradises in very poor countries, like still are in Africa. This to highlight the fact that multiplicity of equilibria is the situation that more represents the actual state of the world, and that is the reason why I decided to focus on it.

In the next section, I will discuss the stability properties of the equilibria, limiting the analysis to a local level. Such a kind of analysis, is then deepened in Sect. 3 by studying the global dynamics of the model. Local analysis is indeed of little relevance when multiplicity of equilibria arises, because it is only able to draw conclusions only on a close neighbourhood of the equilibria, and it is silent about the dynamics in between them.

The first interesting information one can extract from the study of the local dynamics of the equilibria is the occurrence of an eventual poverty trap. From the pictures displayed previously, some equilibria a characterised by low levels of consumption and capital, and some by higher levels of consumption and capital. If more than one equilibria is found to be (saddle) stable, and one provides a lower level of welfare (either because consumption is lower and/or pollution is higher), we may talk about poverty trap, that is to say an equilibria which is socially dominated but from which is difficult to escape.

The second interesting information that will be analysed in the section concerning the global dynamics, is the behaviour of the system in a generic point of the $k - x$ space, which represent the initial conditions, respectively, for capital and pollution. The question I will try to answer is whether the social planner will bring pollution to its irreversibility region or not, starting, as an example, in a neighbourhood of an unstable equilibria or far enough from a stable equilibria.

Local dynamics of the equilibria. The study of the local dynamics of the system around the steady states is usually carried on by linearising the model around them, using a first order taylor expansion. The first order taylor expansion or Jacobian matrix of the system (31)–(34) is therefore:

$$\begin{pmatrix} \dot{\tilde{k}} \\ \dot{\tilde{x}} \\ \dot{\tilde{\lambda}}_1 \\ \dot{\tilde{\lambda}}_2 \end{pmatrix} = \begin{pmatrix} A-\delta & 0 & \frac{1}{\sigma}\lambda_1^{*-\frac{1+\sigma}{\sigma}} & 0 \\ A & -\eta+2\theta x^* & 0 & 0 \\ 0 & 0 & \rho-(A-\delta) & -A \\ 0 & -2\theta\lambda_2^*+B(\gamma-1)x^{*\gamma-2} & 0 & \eta-2\theta x^*+\rho \end{pmatrix} \cdot \begin{pmatrix} \tilde{k} \\ \tilde{x} \\ \tilde{\lambda}_1 \\ \tilde{\lambda}_2 \end{pmatrix}$$

The characteristic polynom of the matrix of coefficient can be written as

$$(\mu^2 - \rho\mu)^2 + (\mu^2 - \rho\mu)z + s \tag{37}$$

with

$$z = (A - \delta)(\rho - (A - \delta)) - (\eta - 2\theta x^*)(\rho + \eta - 2\theta x^*) \tag{38}$$

$$s = (A - \delta)(A - \delta - \rho)(\eta - 2\theta x^*)(\rho + \eta - 2\theta x^*) + \\ + \frac{1}{\sigma}\lambda_1^{*-\frac{1+\sigma}{\sigma}}\left\{A^2\left[-2\theta\lambda_2^* + B(\gamma - 1)x^{*\gamma-2}\right]\right\} \tag{39}$$

Equating the characteristic polynom to zero, and computing the eigenvalues, I get:

$$\mu_{1,2,3,4} = \frac{1}{2}\rho \pm \sqrt{(\rho/2)^2 - \frac{1}{2}z \pm \frac{1}{2}\sqrt{z^2 - 4s}} \tag{40}$$

and the following lemmas hold:

1. If $z < 0$, $0 < s \leq (z/2)^2$ it is a nec. and suff. condition for all μ to be real, 2 positive and two negative.
2. If $s > (z/2)^2$ and $s - (z/2)^2 - \rho^2 \cdot (z/2) > 0$ it is a nec. and suff. condition for all μ to be complex, two with negative real parts and two with positive real parts.
3. If $s < 0$ it is a nec. and suff. condition for one eig. to be negative and either 3 eig. to be positive or one positive and two having positive real parts.
4. If $s > (z/2)^2$ and $s - (z/2)^2 - \rho^2 \cdot (z/2) = 0$ it is a nec. and suff. condition for all μ to be complex and two having zero real part.

It follows from these lemmas that any equilibrium lying on the increasing locus of the marginal rate of decay is saddle-stable ($z < 0$), while the equilibria lying on the decreasing part of the natural rate of decay of pollution are stable if and only if $s > 0$. Since s depends on equilibrium levels of λ_1, λ_2 and x, analytical conditions determining the sign of s cannot be found and therefore we have to rely on numerical simulations.

The next table presents two possible outcomes which are depicted in Fig. 1 above. The first block is about the results obtained using the parameters of the first two pictures (case 1 and 2) and considering a minimum level of consumption mc very low (in particular, $mc = 0.0005$). It follows from those estimation that the model exhibits two stationary and stable solutions out of four, implying that under the set of parameters used, only the equilibria on the increasing locus of the decay function of pollution are stable. However, the second block shows that, changing just one parameter (in particular, bringing ρ from 0.04 to 0.02, which is the set of parameters used in the third and fourth pictures above) the corner equilibria lying on the decreasing locus of the function of the pollution's decay is stable. This means that lowering the level of impatience of the representative citizen, it is better to stay in the (socially) dominated equilibria than deviating. Of course this result hold in a close neighbourhood of the equilibria provided that capital respects the constraint

of being greater than $\underline{k}$. This result can already be an indicator of a non-optimality of irreversible pollution accumulation since if the population has a "low enough" rate of intertemporal preferences, the higher levels of utility they can achieve by increasing capital and consumption are not big enough to compensate the losses due to the growth of pollution.

	E1	E2	E3	E4
Case 1 $mc =$ 0.0005	$k^* = 0.2543$ $x^* = 0.5240$ $\lambda_1^* = 0.1758e + 08$ $\lambda_2^* = -0.1450e + 08$ $\mu_1 = 0.6995$ $\mu_2 = 0.0556$ $\mu_3 = -0.0156$ $\mu_4 = -0.6595$	$k^* = 0.0025$ $x^* = 0.5240$ $\lambda_1^* = 1.8911e + 08$ $\lambda_2^* = -1.5602e + 08$ $\mu_1 = 0.6999$ $\mu_2 = 0.0308$ $\mu_3 = 0.0092$ $\mu_3 = -0.6599$	$k^* = 0.0007$ $x^* = 0.0197$ $\lambda_1^* = 4,000,000$ $\lambda_2^* = -12965.5$ $\mu_1 = -0.0280$ $\mu_2 = -0.66$ $\mu_3 = 0.0680$ $\mu_3 = 0.7$	$k^* = 0.0007$ $x^* = 0.5803$ $\lambda_1^* = 4,000,000$ $\lambda_2^* = -1.22e + 08$ $\mu_1 = 0.0119$ $\mu_2 = -0.66$ $\mu_3 = 0.0281$ $\mu_3 = 0.7$
Case 3 $mc =$ 0.0005	$k^* = 0.0052$ $x^* = 0.5240$ $\lambda_1^* = 0.1758e + 08$ $\lambda_2^* = -0.1450e + 08$ $\mu_1 = 0.6996$ $\mu_2 = 0.0386$ $\mu_3 = -0.0186$ $\mu_4 = -0.6796$	$k^* = 0.0007$ $x^* = 0.0197$ $\lambda_1^* = 8e + 09$ $\lambda_2^* = -129455.1$ $\mu_1 = -0.028$ $\mu_2 = -0.68$ $\mu_3 = 0.048$ $\mu_3 = 0.7$	$k^* = 0.0007$ $x^* = 0.5803$ $\lambda_1^* = 8e + 09$ $\lambda_2^* = -1.17e + 08$ $\mu_1 = -0.008$ $\mu_2 = -0.68$ $\mu_3 = 0.028$ $\mu_3 = 0.7$	

Depending on the parameter's set chosen, this model can predict two types of poverty traps, one characterised by low levels of pollution, and one characterised by high levels of pollution. This outcome is in line with the evidence on the environmental quality in different poor countries. It is indeed not rare that very polluted cities in the world are often situated in poor countries. According to the Time, for instance, the most polluted places in the world are in China and India.[4] The evidence suggests moreover that the cleanest cities in the world[5] are located in developed (rich) countries, suggesting that those places are probably near the interior equilibria characterised by relatively low levels of pollution and high levels of per capita income.

The next section instead studies the existence of an optimal path conducing pollution to its irreversibility region, and later the two solutions are compared through simulation, using a dynamic programming algorithm.

[4]Linfen (China) is the first most polluted city in the world, where the amount of particulate matters in the air is such that it makes the laundry black before it dries, followed by Sukinda (China) where 60% of the drinking water contains hexavalent chromium at levels more than double international standards and Vapi (India), where levels of mercury in the city's groundwater are reportedly 96 times higher than WHO safety levels, and heavy metals are present in the air and the local produce.

[5]Among the cleanest cities in the world we see Calgary (Canada), Honolulu (Hawaii), Helsinki (Finland), and Ottawa (Canada).

2.2 Irreversible Pollution Accumulation

This solution requires that pollution reaches its threshold level. The analysis of this second option available to the planner has a point of non-differentiability (in $x = \bar{x}$) and therefore I use the same approach kept by Tahvonen and Whithagen and I split the problem into two subproblems: the first - or first period problem - where the planner maximises in finite time T the citizen's discounted utility, with the constraint that $x_T = \bar{x}$, and the second period problem, where the maximisation goes from T to infinity, with initial conditions $x_T = \bar{x}$ and k_T equal to the final value of capital in the first period. Of course, in the second period problem the natural decay function for pollution is zero since it has reached the threshold of irreversibility.

The problem can be expressed, then, as follows[6]:

$$\max_{c(t)} W^T = \int_0^T e^{-\rho t}\left[\frac{C(t)^{1-\sigma}-1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right]dt \tag{41}$$

subject to

$$\dot{k}(t) = (A-\delta)k(t) - C(t) \tag{42}$$

$$\dot{x}(t) = Ak(t) - \eta\left(x(t) + \frac{\theta}{\eta}x(t)^2\right) \tag{43}$$

$$k(0) = k_0 \tag{44}$$

$$k(T) \geq \underline{k} \tag{45}$$

$$x(0) = x_0, \qquad x_0 < \bar{x} \tag{46}$$

$$x(T) = \bar{x}, \qquad T < \infty \tag{47}$$

which represents the so-called "first period problem",[7] immediately followed by the "second period problem" that is

$$\max_{c(t)} W_T = \int_T^{\infty} e^{-\rho t}\left[\frac{C(t)^{1-\sigma}-1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right]dt \tag{48}$$

subject to the laws of motion of the two state variables and the initial conditions

$$\dot{k}(t) = Ak(t) - \delta k(t) - C(t) \tag{49}$$

$$\dot{x}(t) = Ak(t) \tag{50}$$

[6]It is worthwhile here to make some clarifications: let T the number of periods the planner chooses to let pollution reach its own threshold of irreversibility. It might be the case that (i) The planner fixes an arbitrary T and set $x(T) = \bar{x}$ or (ii)The planner chooses the optimal T such that $x(T) = \bar{x}$. Both cases are admissible, however, in the second case further optimality conditions are required and are explained in the text.

[7]Transversality conditions are not required in the first period problem.

$$x(T) = \bar{x} \tag{51}$$

$$k(T) = k_T \tag{52}$$

and that is also subject to the following sign and transversality conditions, which are sufficient conditions for maximisation:

$$\lambda_1(t) \geq 0 \tag{53}$$

$$\lim_{t\to\infty} e^{-\rho t}\lambda_1(t)k(t) = 0 \tag{54}$$

$$\lambda_2(t) \leq 0 \tag{55}$$

$$\lim_{t\to\infty} e^{-\rho t}\lambda_2(t)\big(x(t) - \bar{x}\big) = 0 \tag{56}$$

It is possible to prove (proof provided in the appendix) that an optimal path for the first period problem exists, because all the state variables are closed subset of $\mathbb{R}$, and the control $C(t) \in \mathscr{C} \subseteq \mathbb{R}$.

Let us denote the maximised welfare function for the first and second period, respectively, $\hat{W}^T$ and $\hat{W}_T$ for $T < \infty$. The maximised utility function for the whole period is then $W = \hat{W}^T + \hat{W}_T$. If T is considered fixed, nothing has to be added to the problem, otherwise, if the planner wish to chose the optimal T, let's say T^*, the maximum principle requires that in addition to the first order conditions and transversality conditions, also this condition must be satisfied:

$$\mathscr{H}(k^*(T^*), x^*(T^*), c^*(T^*), \lambda_1(T^*), \lambda_2(T^*), T^*) = 0 \tag{57}$$

The existence of an optimal control with free final time is proved in the appendix A, provided we modify the assumptions such that T^* is free to vary in $[T_1, T_2]$ and the theorem is satisfied on the interval $[0, T_2]$. If the planner wishes to maximise the utility by choosing the optimal terminal time T^*, it must be the case that

$$\left.\frac{\partial W}{\partial T}\right|_{T^*} = \left.\frac{\partial \hat{W}^T}{\partial T}\right|_{T^*} + \left.\frac{\partial \hat{W}_T}{\partial T}\right|_{T^*} \tag{58}$$

where

$$\begin{aligned} e^{\rho T}\frac{\partial \hat{W}^T}{\partial T} &= \frac{c^T(T)^{1-\sigma} - 1}{1-\sigma} - \frac{B\bar{x}^\gamma}{\gamma} + \lambda_1^T(T)[(A-\delta)k^T(T) - c^T(T)] + \\ &+ \lambda_2^T(T)[Ak^T(T) - \underbrace{\eta(\bar{x} + \frac{\theta}{\eta}\bar{x}^2)}_{=0}] \end{aligned} \tag{59}$$

$$\begin{aligned} -e^{\rho T}\frac{\partial \hat{W}_T}{\partial T} &= \frac{c_T(T)^{1-\sigma} - 1}{1-\sigma} - \frac{B\bar{x}^\gamma}{\gamma} + \lambda_{1T}(T)[(A-\delta)k_T(T) - c_T(T)] + \\ &+ \lambda_{2T}(T)[Ak_T(T)] \end{aligned} \tag{60}$$

Equation 58 is verified when $C_T(T) = C^T(T)$ and $k_T(T) = k^T(T)$, i.e. they are continuous functions at T^*. If instead it is not optimal to reach $\bar{x}$ in finite time, it must be the case that

$$\lim_{T\to\infty} \sup \frac{\partial W}{\partial T} = \lim_{T\to\infty} \sup \frac{\partial \hat{W}^T}{\partial T} + \lim_{T\to\infty} \sup \frac{\partial \hat{W}_T}{\partial T} \geq 0 \tag{61}$$

so it is necessary that W does not decrease when T increases without limit.

For what concerns the first period problem, define the current value Hamiltonian associated to the problem (41)–(47) as

$$\mathscr{H}(t, k(t), x(t), c(t), \Lambda; \Theta) \stackrel{def}{=} \cdot \left[\frac{C(t)^{1-\sigma} - 1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right] +$$
$$+\lambda_1(t)\Big[(A-\delta)k(t) - c(t)\Big] + \lambda_2(t)\Big[Ak(t) - \eta\big(x(t) + \frac{\theta}{\eta}x(t)^2\big)\Big] \tag{62}$$

where Λ is the set of shadow prices and Θ represents the set of exogenous parameters of the model where, as before, $\Theta = \{A, B, \sigma, \rho, \delta, \eta, \theta, \gamma, mc\}$.
The necessary first order conditions are:

$$\frac{\partial \mathscr{H}}{\partial c} = 0 \quad \Rightarrow \quad \lambda_1 = C^{-\sigma} \tag{63}$$
$$\frac{\partial \mathscr{H}}{\partial k} = \rho\lambda_1 - \dot{\lambda}_1 \quad \Rightarrow \quad \dot{\lambda}_1 = \lambda_1(\rho + \delta - A) - \lambda_2 A \tag{64}$$
$$\frac{\partial \mathscr{H}}{\partial x} = \rho\lambda_2 - \dot{\lambda}_2 \quad \Rightarrow \quad \dot{\lambda}_2 = \lambda_2 \cdot [\rho + \eta\big(1 - \frac{2\theta}{\eta}x\big)] + Bx^{\gamma-1} \tag{65}$$

so the following system of four differential equations represents the conditions any optimal path has to obey:

$$\dot{k} = (A-\delta)k - \lambda_1^{-\frac{1}{\sigma}} \tag{66}$$
$$\dot{x} = Ak - \eta x + \theta x^2 \tag{67}$$
$$\dot{\lambda}_1 = \lambda_1(\rho - (A-\delta)) - \lambda_2 A \tag{68}$$
$$\dot{\lambda}_2 = \lambda_2(\eta - 2\theta x + \rho) + Bx^{\gamma-1} \tag{69}$$
$$k(0) = k_0 \tag{70}$$
$$k(T) \geq \underline{k} \tag{71}$$
$$x(0) = x_0, \qquad x_0 < \bar{x} \tag{72}$$
$$x(T) = \bar{x}, \qquad T < \infty \tag{73}$$

with Eqs. 68 and 69 representing the Euler equations.

For the second period problem, define the current value Hamiltonian associated to the problem (48)–(52) as:

$$\mathscr{H}(t, k(t), x(t), c(t), \Lambda; \Theta) \overset{def}{=} \lambda_0\left[\frac{C(t)^{1-\sigma}-1}{1-\sigma} - \frac{Bx(t)^{\gamma}}{\gamma}\right] + \\ +\lambda_1(t)\Big[Ak(t) - \delta k(t) - C(t)\Big] + \lambda_2(t)\cdot Ak(t) \tag{74}$$

where Λ is the set of shadow prices and, as before $\Lambda = \{\lambda_0(t), \lambda_1(t), \lambda_2(t)\}$ with λ_1 and λ_2 representing, respectively, the shadow prices of capital and pollution and Θ represents the set of exogenous parameters of the model where, as before, $\Theta = \{A, B, \sigma, \rho, \delta, \eta, \theta, \gamma, mc\}$.

The maximum principle asserts that there exists a λ_0 and a continuous and piecewise continuously differentiable functions $\lambda_1(t)$ and $\lambda_2(t)$, such that for all t

$$(\lambda_0, \lambda_1(t), \lambda_2(t)) \neq (0, 0, 0) \tag{75}$$

$$\mathscr{H}(t, k^*(t), x^*(t), c^*(t), \Lambda; \Theta) \geq \mathscr{H}(t, k^*(t), x^*(t), c(t), \Lambda; \Theta) \qquad \forall t \tag{76}$$

Moreover,[8]

$$\frac{\partial \mathscr{H}}{\partial c} = 0 \quad \Rightarrow \quad \lambda_1 = C^{-\sigma} \tag{77}$$

$$\frac{\partial \mathscr{H}}{\partial k} = \rho\lambda_1 - \dot{\lambda}_1 \quad \Rightarrow \quad \dot{\lambda}_1 = \lambda_1(\rho + \delta - A) - \lambda_2 A \tag{78}$$

$$\frac{\partial \mathscr{H}}{\partial x} = \rho\lambda_2 - \dot{\lambda}_2 \quad \Rightarrow \quad \dot{\lambda}_2 = \lambda_2 \cdot \rho + Bx^{\gamma-1} \tag{79}$$

$$\lambda_0 = 1 \quad \text{or} \quad \lambda_0 = 0 \tag{80}$$

Since the terminal conditions for capital and pollution as time approaches infinity are left free, it follows that $\lim_{t\to\infty} \lambda_1(t) = 0$ and $\lim_{t\to\infty} \lambda_2(t) = 0$ so $\lambda_0 = 1$. Finally, (51) and (52) have to be satisfied.

Rearranging Eq. 77 we get an expression for consumption in terms of the shadow price of capital:

$$C = \lambda_1^{-\frac{1}{\sigma}} \tag{81}$$

so the economic system can be represented by the following four differential equations

$$\dot{k} = (A - \delta)k - \lambda_1^{-\frac{1}{\sigma}} \tag{82}$$

$$\dot{x} = Ak \tag{83}$$

$$\dot{\lambda}_1 = \lambda_1(\rho + \delta - A) - \lambda_2 A \tag{84}$$

$$\dot{\lambda}_2 = \lambda_2\rho + Bx^{\gamma-1} \tag{85}$$

[8]For notational simplicity, in the following I will use interchangeably the generic variable z instead of $z(t)$ whenever this does not constitute ambiguity.

From Eq. 83 is straightforward to see that in order to keep pollution stable through time ($\dot{x} = 0$) it is necessary to keep capital nil. This means that no production nor consumption can occur in steady state, and since for hypothesis the model guarantees a minimum level of consumption, implying also a strictly positive capital and production, no stationary solution can be found in this second period problem. In the long run, the optimal path will converge to a consumption level equal to the subsistence level mc, with a capital level constant and equal to $\underline{k}$. This level of capital is such that it produces a level of income which sustains a minimum level of consumption and an investment level which is equal to the depreciation of capital.

The existence of a balanced growth path for capital, pollution and consumption can be reasonably excluded because this would imply a constant and equal rate of growth for all the variables involved, consumption, pollution and capital. Since the marginal utility from consumption is an increasing and concave function of consumption, and the marginal disutility from pollution is an increasing and convex function of pollution (so it grows at a rate that is greater than the rate of growth of the marginal utility from consumption), there will be a point on time $t' \in [T, \infty)$ where an additional unit of pollution will produce a disutility higher than the utility produced by an additional unit of consumption. At this point in time, the optimal path will predict a consumption level equal to mc, production equal to $\underline{y}$ and a minimum level of capital $\underline{k}$ that is necessary to guarantee a level of investments that covers the depreciation, and a subsistence level of consumption. Pollution, from t' onward, will have an instantaneous variation $\dot{x} = A\underline{k}$, while the variation of capital and consumption will be nil.

2.3 Paths Comparison

The model does not allow to say which of the paths gives higher utility, so direct comparison is necessary. In particular, we are interested to see whether an irreversible solution may provide an higher level of discounted utility than an irreversible one. But this requires first of all the computation of W^{∞} and $W = \hat{W}^T + \hat{W}_T$. The analysis presented in paragraph 2.2.1 is only partial, because it is just able to say something about the stability of the two steady states and their associated level of welfare W^{∞} if the system is in equilibrium and there are no shocks able to carry on the system far away from them, but is is completely unable to say anything about the behaviour of the system in between of the two fixed points, or in any point of the $x - k$ plane. In order to say something about the behaviour of this economy far away from the equilibria, global analysis is needed. In the next section, I will use an algorithm of dynamic programming to carry on this analysis.

As it was previously anticipated, the problem presented here has the peculiarity of having, for each initial condition and in finite time horizon, two simultaneous optimal paths which respect the first order conditions. In accordance to the possibilities available to the planner, it may be optimal either to increase utility by reducing pollution or, viceversa, by increasing consumption. The first choice takes the system

in a path which is converging to the saddle stable equilibria introduced in Sect. 2.1 (so in this case a reversible solution is optimal), and the other choice takes the system toward the irreversibility threshold for pollution. Whether it is optimal one or the other, is a question I will try to answer below.

3 Global Analysis

In this section, I am interested to see whether - in case of multiple equilibria - the system can, starting from a neighbourhood of the unstable equilibria, recover and converge to the socially optimum steady state. Due to the lack of closed form solution of this dynamic model, I need to use computational methods. I use the convenient approach of dynamic programming, which provides the value function and the control variable in feedback form. This allows to find the global dynamics of the state space in the region restricted by arbitrary values of capital and pollution, using a fixed grid size technique.

3.1 Discretisation

The first step to do that is to discretize the model identified by Eqs. 16–18

$$\max_{c_t \in \mathscr{C}_t} U_t = \sum_{t=0}^{\infty} \beta^t \left[\frac{c_t^{1-\sigma} - 1}{1-\sigma} - B \frac{x_t^{\gamma}}{\gamma} \right] \tag{86}$$

subject to

$$x_{t+1} = Ak_t - x_t(\eta - \theta x_t - 1) \tag{87}$$
$$k_{t+1} = (A - \delta + 1)k_t - c_t \tag{88}$$
$$\beta = (1 - \rho) \tag{89}$$
$$k_0 = k \tag{90}$$
$$x_0 = x \tag{91}$$

Here $\mathscr{C}_t$ denotes the set of discrete control sequences $C = (C_1, C_2, ...)$ for $C_i \in \mathscr{C}$ The optimal value function V is the unique solution of the discrete Hamilton-Jacobi-Bellman's equation

$$V(k, x) = \max_{c \in \mathscr{C}} \left\{ u_t(k_t, x_t, C_t) + \beta V(k_{t+1}, x_{t+1}) \right\} \tag{92}$$

with

$$u_t(k_t, x_t, C_t) = \left[\frac{c_t^{1-\sigma} - 1}{1-\sigma} - B\frac{x_t^{\gamma}}{\gamma} \right] \tag{93}$$

If I define the dynamic programming operator T by

$$T(V)(k, x) = \max_{C \in \mathscr{C}} \left\{ u_t(k_t, x_t, C_t) + \beta V(k_{t+1}, x_{t+1}) \right\} \tag{94}$$

then V can be characterised as the unique solution of the fixed point equation

$$V(k, x) = T(V)(k, x) \qquad \text{for all } x, k \in \mathscr{R}^n \tag{95}$$

3.2 Results

The study of dynamic decision models with multiple equilibria is intricate. Multiple equilibria can arise in models with non-concave pay-off functions, externalities and increasing returns. Recently multiple equilibria have been found also in concave economies (for a survey on models with multiple equilibria, see Deissenberg et al. [6]). In terms of dynamics, multiple equilibria are difficult to analyse, since the domain of attraction might not coincide with the stable and unstable equilibria, and multiple optimal paths may exist as well. In the context of my model, multiple (non-trivial) equilibria arise from some parameter constellations. In the following, I will consider only a set of parameters which gives multiple steady states, because I believe this case is the most interesting from a policy point of view. Consider the following parameter set:

$$\begin{aligned} B &= 10,000,000 \\ A &= 0.8 \\ \rho &= 0.04 \\ \theta &= 0.05 \\ \eta &= 0.03 \\ \sigma &= 3 \\ \gamma &= 3 \\ \delta &= 0.1 \\ \beta &= 1 - \rho = 0.96 \end{aligned}$$

Those parameters yield the following numerical solution for the two (non-trivial) steady states:

Variable	Equilibrium 1	Equilibrium 2
k^*	0.5494e-2	0.2489e-2
x^*	0.2543	0.5240
λ_1^*	0.1758e+08	1.8911e+08
λ_2^*	-0.1450e+08	-1.5602e+08

The eigenvalues of the first equilibria are $\mu_1 = 0.6995$, $\mu_2 = 0.0556$, $\mu_3 = -0.0156$ and $\mu_4 =$ -0.6595 and for the second are $\mu_1 = 0.6999$, $\mu_2 = 0.0308$, $\mu_3 = 0.0092$ and $\mu_4 = -0.6599$. This information allows us to say only that the first equilibria is saddle stable (however, this conclusion holds only locally and what happens between the two steady states is a black box), and that second is unstable. Nothing can be said about the direction of the instability of this latter equilibria. In other words, from the local analysis nothing can be inferred about whether - starting from initial conditions close to the unstable equilibria - the system will converge to the stable (and pareto dominant) equilibria or not. To this purpose, I studied the global dynamics of this system using a dynamic programming algorithm. Dynamic programming allows to draw the phase diagram of the system in terms of the states variables and it is a convenient tool to study the global dynamics in case of multiplicity of equilibria. The algorithm is described in detail in appendix, and results are depicted in Fig. 2.

The first thing is to check first of all if the system will converge to the socially dominant equilibria or not, starting in proximity of the unstable one. Numerical simulations show that, for example, assuming a fixed plan horizon of 50 periods, there exist two optimality candidates: one path that brings pollution toward the irreversibility region and the other one that converges to the saddle stable (and socially dominant) equilibria. Basically, an efficiently managed economy may choose to achieve the objective of maximising the utility function by means of two instruments: (i) increasing consumption or (ii) reducing pollution. The first policy implies that the consumption profile of the first periods is left low, capital is allow to increase at a very fast rate, and so also consumption in the subsequent periods. The second policy, viceversa, is described by an high level of consumption in the first period (aimed at reducing the level of capital, responsible for the production of pollution), and a low profile (although increasing) of consumption in subsequent periods. The choice between these two paths cannot be made a priori and the computation of the utility's present value is needed.

Figures 2 and 3 show the phase diagrams in terms of the state variables and the behaviour of the control variable for the two different paths. Numerical simulations show that the utility's present value for the first path (the path diverging towards the irreversibility threshold of pollution, represented in Fig. 2) is equal to $-1.1427e + 07$, against a present value of $-1.4796e + 07$ for the second path in Fig. 3, the path converging to the saddle stable equilibria. It is therefore worth increasing capital and consumption up or close to the irreversibility threshold of pollution, if the time horizon is sufficiently low.

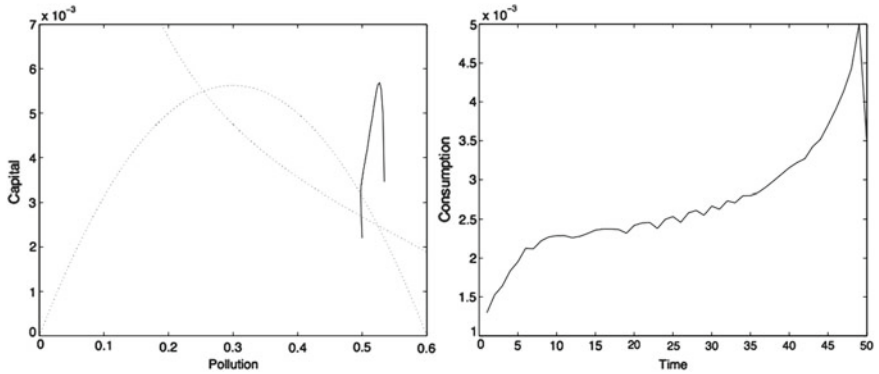


Fig. 2 Divergent path, T = 50, $k_0 = 0.0025$, $x_0 = 0.5$

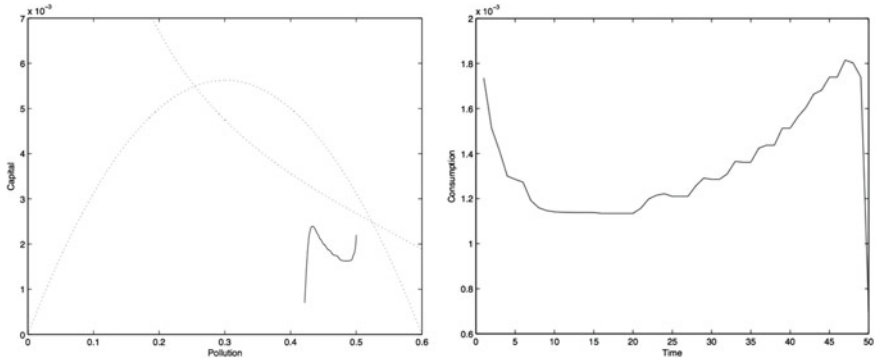


Fig. 3 Convergent path, T = 50, $k_0 = 0.0025$, $x_0 = 0.5$

Things seem different, however, when the plan horizon is longer. As T grows, hypothetically to infinity, numerical simulations suggest that the optimal path is no longer to bring pollution close or up to the irreversibility region, as it is clearly highlighted in the figures below. Those pictures indeed show a tendency of the system to converge to the stable and socially optimal interior equilibria.[9] This may be explained by the fact that the utility of having high levels of consumption for limited amounts of periods followed by minimum levels of consumption and increasing levels of pollution for infinite periods is definitely worse than having moderately high levels of consumption and low levels of pollution forever, especially if the intertemporal rate of preferences is not too high.

As it is possible to see in Fig. 4, for $T = 200$, the tendency is to keep capital at a level which allows both strictly positive consumption and a reduction of pollution, except for dramatically increase capital, consumption and pollution during the last periods (but this is due to the fact that we are dealing with a routine that in order to be ran has to set a finite time horizon).

[9]The minimum consumption level is assumed to be set at very low levels such that the solution never hits it.

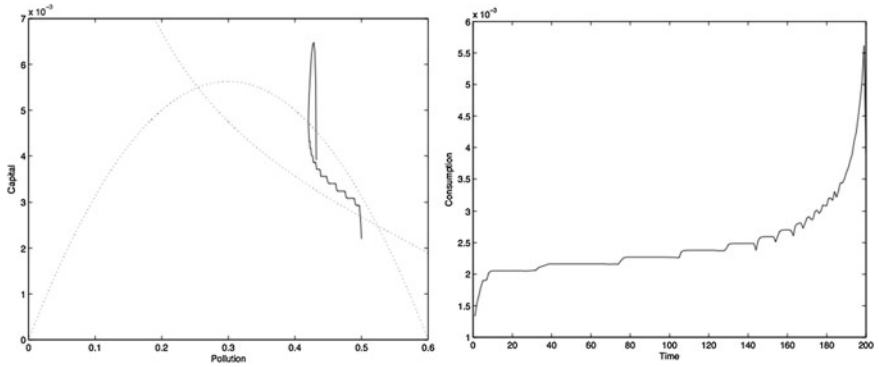


Fig. 4 Optimal plan, T = 200, $\rho = 0.04$, $k_0 = 0.0025$, $x_0 = 0.5$

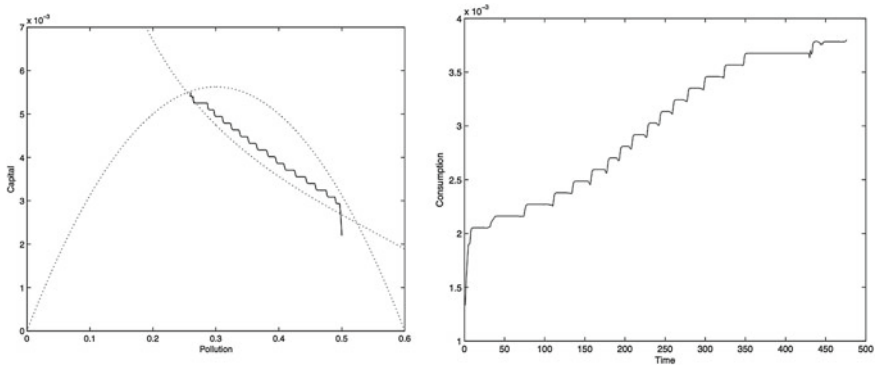


Fig. 5 Optimal plan, T = ∞, $\rho = 0.04$, $k_0 = 0.0025$, $x_0 = 0.5$

So, two identical countries may choose different environmental policies only if they differ in the choice of their planning horizons. Governments who have short term objectives will choose paths which imply a growing stock of pollution through time, whilst governments with longer horizons will choose paths which imply a decrease of pollution and a slower increase in consumption. This convergence to the interior and stable equilibria is found irrespective of the set of initial conditions.

Figure 7 displays a path leading pollution to reach its irreversibility threshold. As predicted in the previous section, after pollution has become irreversible, it is optimal for the planner to let capital and consumption to reach their "survival" levels set by $\underline{k}$ and mc, respectively. The picture shows also that pollution grows steadily through time.

In order to compare the two different environmental policies, it is necessary to compute the present value of all the flows of utility provided in each period by the two paths. The optimal path depicted in Fig. 5 provides a present value of utility equal to $-5.93e + 07$, whilst the path depicted in Fig. 7 provides $-2.09e + 08$, with a negative difference of $-1.50e + 08$.

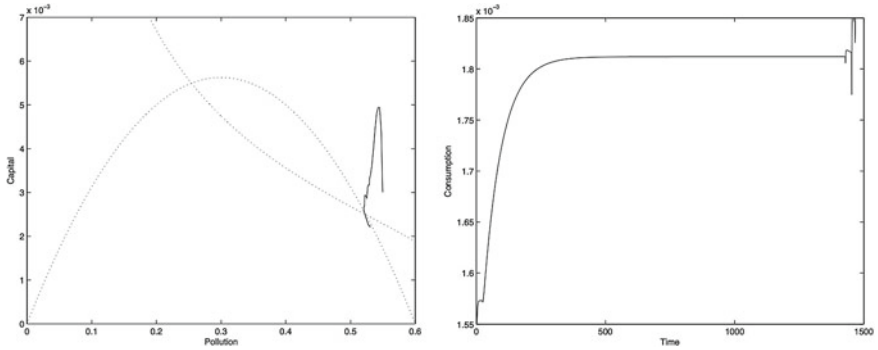


Fig. 6 Optimal plan, T = 1500, $\rho = 0.04$, $k_0 = 0.0022$, $x_0 = 0.5$

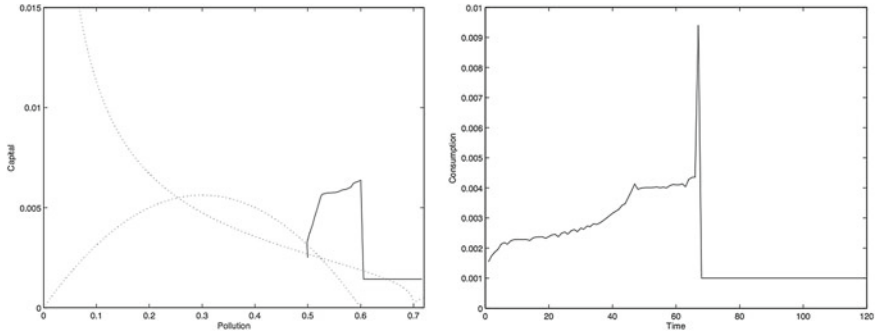


Fig. 7 Optimal plan, T = ∞, $\rho = 0.04$, $k_0 = 0.0025$, $x_0 = 0.5$, $mc = 0.001$

Contrary to Tahvonen and Withagen's result, according to which optimality or not of irreversible pollution accumulation depends on the initial condition for pollution (being optimal when the initial condition for pollution is higher than the level determined by the unstable equilibria), my numerical simulations show a different story. Figure 6 provides a clear example. Initial condition for pollution is 0.53, which is higher than 0.524 characterising the second equilibria. The time span necessary to get into what they call "domain of attraction of the saddle stable equilibria" is however very high, and increases as the initial pollution level increases. Consumption also grows steadily but slowly in proximity of the equilibria: unfortunately the accuracy of the picture is not enough to make it evident. What makes the difference between their model and my model is not the set of initial condition (which is also a poor explanation of the reasons why a country should prefer irreversibility), but the fact that the marginal utility of consumption when consumption is zero, is nil. This is an implication of the fact that they deal with a local pollution problem and not with a global one. Being zero the marginal utility from consumption when there is no consumption at all implies that the population can move elsewhere to satisfy their needs. In my model it is not possible, and there survival (and consumption, although at minimum levels) is always preferred to an additional unit of pollution (Fig. 7).

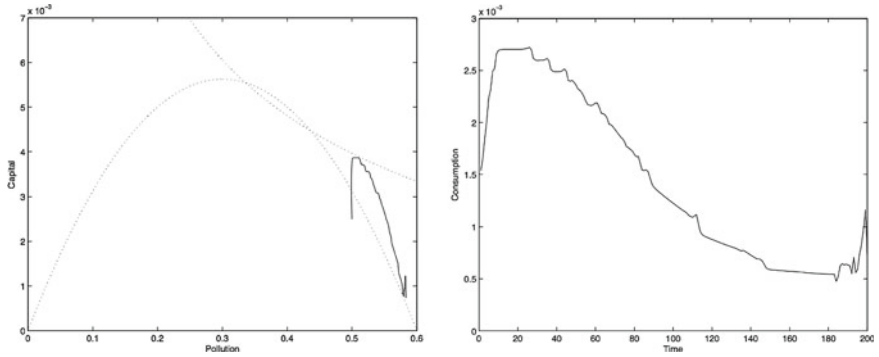


Fig. 8 Optimal plan, T = 200, $\rho = 0.09$, $k_0 = 0.0025$, $x_0 = 0.5$

The simulations therefore are clear in highlighting the fact that irreversibility is never optimal, if the planning horizon is infinite, and no matters the initial level of pollution. Different environmental policies can therefore be explained only by different planning horizon of the governments, all other parameters constant.

3.3 The Effect of ρ on the Global Dynamics of the System

The representative household's level of impatience, represented by ρ, may play a crucial role in determining the environmental policy chosen by the planner. The more impatient the people are, the more probable is a policy which implies a growing stock of pollution through time. The effects are someway similar to a shortening of the time horizon, and this is confirmed by the simulations. Figure 8 represents the global dynamics of the system assuming a time horizon of 200 periods, and an intertemporal rate of preferences equal to 0.09. The only parameter that distinguishes Fig. 8 from Fig. 4 is the level of ρ, but as it is possible to see the dynamics is dramatically different. Convergency to the saddle stable equilibria is harder to find, since the high discount rate and the finite horizon make worth for the planner choosing a path which implies an high growth rate of consumption for the first periods, which are valued more than future ones, especially because pollution - compared to consumption, grows at much slower rate.

4 Conclusion

This paper contributes in the debate about the necessity of a unified global environmental policy kept by all the nations. With a theoretical model of economic growth with pollution accumulation and an endogenous function for the natural decay of pollution, I show with numerical simulation that in an efficiently planned economy, with infinite time horizon plan, irreversible pollution accumulation cannot be an

optimal policy. Optimality of such a policy can occur only if the plan of government is short-minded. Since utility depends on both consumption and the level of pollution, in principle the planner can choose to achieve the maximum level of welfare by, alternatively, increasing consumption or reducing pollution. The second strategy pays in the long run, while the first in short run.

It would be important to keep in mind that although each nation decides on its own its environmental policy and it is free to join international agreement, we live in the same planet and if all the countries would be one with the priority to safeguard life, they would not engage in such production of pollution, in each form. So, despite incentives are not enough to give up opportunities to grow in the short run, it would be useful to ask whether such individual policy can be consistent with individual long terms goal. Each country may decide to pollute a lake if there are others from which he can extract utility, but what happens when all the lakes are polluted?

Unfortunately, populist policies and the fact that politicians stay in power for few years and needs to be reelected can - someway - affect the environmental policies undertaken by the countries. Special interests overcome in importance other issues which are in general considered of marginal relevance, like the environment, because a policy undertaken by a single country can only marginally affect it, especially when it deals with global problems.

So, a deep study of the incentives taking a country to engage in global emission's reduction is needed and it can be part of future research.

5 Appendix

A. Proof of the Existence of an Optimal Path for the First Period Problem - T Fixed

To prove the existence of an optimal path for the first period problem when the final time T is fixed and the pollution at T is equal to its threshold, I use the Filippov - Cesari theorem of existence of an optimal control (Seierstad and Sydsæter [13], p. 132) which requires the convexity of the set

$$N(k, x, \mathscr{C}, t) = \left\{ \frac{c^{1-\sigma} - 1}{1-\sigma} - \frac{Bx^{\gamma}}{\gamma} + \omega, (A-\delta)k - c, Ak - \eta\left(x + \frac{\theta}{\eta}x^2\right) \right\} \tag{96}$$

where $\mathscr{C} \subseteq \mathbb{R}$ represents the set of all admissible controls, and $\omega \leq 0$. The theorem states: Consider the standard optimal control problem (41)–(47). Assume that:

- There exists an admissible triple $(k(t), x(t), c(t))$.
- $N(k, x, \mathscr{C}, t)$ is convex for each (k, x, t).
- $\mathscr{C}$ is closed and bounded.
- There exists two numbers $\bar{k}$ and $\bar{x}$ such that $\|k(t)\| \leq \bar{k}$ and $\|x(t)\| \leq \bar{x}$ for all $t \in [0, T]$ and all admissible pairs $(k(t), x(t), c(t))$.

Then, there exists an optimal pair $(k^*(t), x^*(t), c^*(t))$ (with $c^*(t)$ measurable). In order to prove the convexity of the set in (96), let us keep $(k(t), x(t), t)$ fixed, so $k(t) = K$ and $x(t) = X$. Let y_1, y_2, y_3 three arbitrary points in $N(K, X, \mathscr{C}, t)$, i.e.

$$y1 = \left\{\left(\frac{c_1^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \omega_1, (A-\delta)K - c_1, AK - \eta\left(X - \frac{\theta}{\eta}X^2\right)\right\}$$

$$y2 = \left\{\left(\frac{c_2^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \omega_2, (A-\delta)K - c_2, AK - \eta\left(X - \frac{\theta}{\eta}X^2\right)\right\}$$

$$y3 = \left\{\left(\frac{c_3^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \omega_3, (A-\delta)K - c_3, AK - \eta\left(X - \frac{\theta}{\eta}X^2\right)\right\}$$

for some $\omega_1, \omega_2, \omega_3 \leq 0$ and $c_1, c_2, c_3 \in \mathscr{C}$. Let λ_1 and λ_2 two positive constants such that $\lambda_1 + \lambda_2 \leq 1$. I need to prove that $y_4 = \lambda_1 y_1 + \lambda_2 y_2 + (1-\lambda_1-\lambda_2)y_3 \in N(K, X, \mathscr{C}, t)$. Put $\lambda_1 y_1 + \lambda_2 y_2 + (1-\lambda_1-\lambda_2)y_3 = (z_1, z_2, z_3)$.

The first component z_1 is:

$$\begin{aligned} z_1 &= \lambda_1\left(\frac{c_1^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \lambda_1\omega_1 + \\ &+ \lambda_2\left(\frac{c_2^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \lambda_2\omega_2 + \\ &+ (1-\lambda_1-\lambda_2)\left(\frac{c_3^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + (1-\lambda_1-\lambda_2)\omega_3 \qquad (97) \\ &= \left\{\lambda_1\frac{c_1^{1-\sigma}-1}{1-\sigma} + \lambda_2\frac{c_2^{1-\sigma}-1}{1-\sigma} + (1-\lambda_1-\lambda_2)\frac{c_3^{1-\sigma}-1}{1-\sigma}\right\}e^{-\rho t} + \\ &- \frac{BX^\gamma}{\gamma}e^{-\rho t} + \lambda_1\omega_1 + \lambda_2\omega_2 + (1-\lambda_1-\lambda_2)\omega_3 \qquad (98) \end{aligned}$$

Since it is known that W^T is concave in c, so $\frac{\partial^2 W^T}{\partial c^2} \leq 0$, we have

$$\begin{aligned} &\lambda_1\frac{c_1^{1-\sigma}-1}{1-\sigma} + \lambda_2\frac{c_2^{1-\sigma}-1}{1-\sigma} + (1-\lambda_1-\lambda_2)\frac{c_3^{1-\sigma}-1}{1-\sigma} \\ &\leq \frac{\left[\lambda_1 c_1 + \lambda_2 c_2 + (1-\lambda_1-\lambda_2)c_3\right]^{1-\sigma}-1}{1-\sigma} \\ &= \frac{c_4^{1-\sigma}-1}{1-\sigma} \end{aligned}$$

with $c_4 = \lambda_1 c_1 + \lambda_2 c_2 + (1-\lambda_1-\lambda_2)c_3$. Then, $c_4 \in \mathscr{C}$. Using this result, from the last inequality we see that

$$z_1 \leq \left(\frac{c_4^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t} + \lambda_1\omega_1 + \lambda_2\omega_2 + (1-\lambda_1-\lambda_2)\omega_3 \qquad (99)$$

Define $\omega_4 = z_1 - \left(\frac{c_4^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\right)e^{-\rho t}$. Then, from (99),

$$\omega_4 \leq \lambda_1\omega_1 + \lambda_2\omega_2 + (1-\lambda_1-\lambda_2)\omega_3 \leq 0 \qquad \text{since } \omega_1, \omega_2, \omega_3 \leq 0$$

The second and third components, z_2 and z_3 are found similarly to the first:

$$\begin{aligned} z_2 &= \lambda_1[(A-\delta)K - c_1] + \lambda_2[(A-\delta)K - c_2] + (1-\lambda_1-\lambda_2)[(A-\delta)K - c_3] \\ &= (A-\delta)K - (\lambda_1 c_1 + \lambda_2 c_2 + (1-\lambda_1-\lambda_2)c_3) \\ &= (A-\delta)K - c_4 \end{aligned}$$

$$\begin{aligned} z_3 &= \lambda_1\big[AK - \eta\big(X - \frac{\theta}{\eta}X^2\big)\big] + \lambda_2\big[AK - \eta\big(X - \frac{\theta}{\eta}X^2\big)\big] + \\ &\quad + (1-\lambda_1-\lambda_2)\big[AK - \eta\big(X - \frac{\theta}{\eta}X^2\big)\big] \\ &= AK - \eta\big(X - \frac{\theta}{\eta}X^2\big) \end{aligned}$$

Piecing all this together, we see that we have found a $c_4 \in \mathscr{C}$ and a $\omega_4 \leq 0$ such that $\lambda_1 y_1 + \lambda_2 y_2 + (1-\lambda_1-\lambda_2)y_3 = \big\{\big(\frac{c_4^{1-\sigma}-1}{1-\sigma} - \frac{BX^\gamma}{\gamma}\big)e^{-\rho t} + \omega_4, (A-\delta)K - c_4, AK - \eta\big(X - \frac{\theta}{\eta}X^2\big)\big\}$. Hence, $\lambda_1 y_1 + \lambda_2 y_2 + (1-\lambda_1-\lambda_2)y_3 \in N(K, X, \mathscr{C}, t)$ and thus $N(K, X, \mathscr{C}, t)$ is convex.

B. The Dynamic Programming Algorithm

The algorithm approximates the solution on a grid Γ covering a compact subset Ω of the state space. I pick a reasonable set Ω and consider only trajectories which remain in Ω in all future times. I assume that for any point $(k, x) \in \Omega$ there exists at least one control value c such that $(k_{t+1}, x_{t+1}) \in \Omega$ holds. Denoting the nodes of the grid Γ by (k^i, x^j), $i = 1, \ldots, n$ and $j = 1, \ldots, m$, the approximation V^Γ satisfy

$$V^\Gamma(k^i, x^j) = T(V^\Gamma)(k^i, x^j) \tag{100}$$

for all nodes (k^i, x^j) of the grid, where the value of V^Γ for points (k, x) which are not grid points (these are needed for the evaluation of T) is determined by bilinear interpolation. Basically, the standard computational algorithm that is used here can be summarised as follows (cite larson):

1. The first step is to set up a grid for the state variables. Each level of capital k and each level of pollution x are quantised, respectively, to N_k and N_x equidistant levels, from 0 to, respectively, $\bar{k}$ and $\bar{x}$. In total, then, the grid points for the state variables are $N_k \cdot N_x$. The control variable c is quantised to N_c equidistant levels, from 0 to $\bar{c}$.
2. For each point in the grid $(k(i), x(j))$, $i = 1, \ldots, N_k$ and $j = 1, \ldots, N_x$, each control $c(h)$, $h = 1, \ldots, N_c$ is applied, and the next state is computed according to the formulas given by Eqs. 87 and 88. Let us call the next-state value of k and

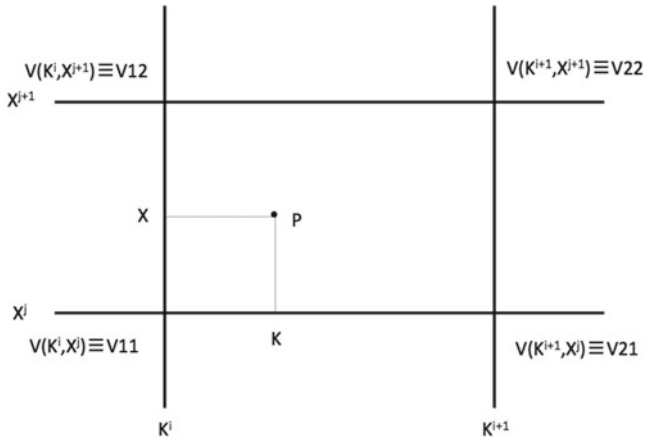
x, respectively, $k1$ and $x1$. Notice that $k1$ and $x1$ are tri-dimensional matrices whose generic element is represented by $k1(i, j, h)$ and $x1(i, j, h)$ and whose dimensions are $N_k \cdot N_x \cdot N_c$. Furthermore, the elements of $k1$ and $x1$ are, in general, not grid points. I then check whether each element of $k1 \in [0, \bar{k}]$ and $x1 \in [0, \bar{x}]$. If they do not belong to those intervals, their values are replaced with "missing".

3. Define the number of periods T, and set up an index $l = 1$. Evaluate k_0 and x_0.
4. The procedure is backward. At the final time T, citizens consume what is left in terms of capital, so $c_T = k_T$ irrespective to the value of x. So, for each point in the grid $(k(i), x(j))$, $i = 1, \ldots, N_k$ and $j = 1, \ldots, N_x$ I compute the value function at time T which is nothing but

$$V^{\Gamma}(k(i), x(j), T) = \frac{k(i)^{(1-\sigma)} - 1}{1 - \sigma} - B\frac{x(j)^{\gamma}}{\gamma} \qquad i = 1, \ldots, N_k, \; j = 1, \ldots, N_x \tag{101}$$

I then store in memory $V^{\Gamma}(k(i), x(j), T)$ and $c(i, j, T) = k(i)$ constant across the j and the T-dimensions.

5. At time $T - l$, for each $i = 1, \ldots, N_k$, $j = 1, \ldots, N_x$ and $h = 1, \ldots, N_c$ I compute the next-period value function $V1(k1(i, j, h), x1(i, j, h))$ interpolating the existing values of $V^{\Gamma}(k(i), x(j), T - l + 1)$ stored in memory. Of course, if either $x1(i, j, h)$ or $k1(i, j, h)$ (or both) are "missing values", also $V1(k1(i, j, h), x1(i, j, h))$ will be "missing". I need to interpolate those values because in general $k1(i, j, h)$ and $x1(i, j, h)$ are not grid points, and I know the value of $V^{\Gamma}(k(i), x(j), T - l + 1)$ only for grid points. Notice that $V1(k1(i, j, h), x1(i, j, h))$ is a tri-dimensional matrix whose dimensions are $N_k \cdot N_x \cdot N_c$. The procedure is the following: the fact that in general $k1(i, j, h)$ and $x1(i, j, h)$, $i = 1, \ldots, N_k$, $j = 1, \ldots, N_x$ and $h = 1, \ldots, N_c$ do not lie on the grid means that I am in the situation in which I have to compute the value function knowing its approximation on four equidistant points around it. Graphically,



I want to approximate a function on the point $P = (k, x)$ that represents my next period values of the state variable once the control is applied, knowing the value function in the points $V11$, $V12$, $V21$ and $V22$. The function in P is computed according to the following formula:

$$\begin{aligned} V(P) = & \frac{V11}{(k^{i+1} - k^i)(x^{j+1} - x^j)} \cdot (k^{i+1} - k)(x^{j+1} - x) \\ & \frac{V21}{(k^{i+1} - k^i)(x^{j+1} - x^j)} \cdot (k - k^i)(x^{j+1} - x) \\ & \frac{V12}{(k^{i+1} - k^i)(x^{j+1} - x^j)} \cdot (k^{i+1} - k)(x - x^j) \\ & \frac{V22}{(k^{i+1} - k^i)(x^{j+1} - x^j)} \cdot (k - k^i)(x - x^i) \end{aligned} \tag{102}$$

6. For each $i = 1, \ldots, N_k$, $j = 1, \ldots, N_x$ and $h = 1, \ldots, N_c$ I compute

$$\begin{aligned} V'(k(i), x(j), c(h), T - l) = & \frac{c(h)^{1-\sigma} - 1}{1 - \sigma} - B\frac{x(j)^{\gamma}}{\gamma} + \\ & \beta V1(k1(i, j, h), x1(i, j, h)) \end{aligned} \tag{103}$$

After that, for each $i = 1, \ldots, N_k$, $j = 1, \ldots, N_x$ I chose the maximum variable over the h-dimension (control), by direct comparison. Those values are stored $c(i, j, T - l)$ and $V^{\Gamma}(k(i), x(j), T - l)$

7. The value of l takes $l + 1$.
8. I check whether l is equal to T. If it is not, I go back to point 5. If $l = T$, then
9. Define three vectors $k^*(t)$, $x^*(t)$ and $c^*(t)$, $t = 1, \ldots, T$ which represent the optimal trajectories of capital, pollution and consumption starting from the initial conditions x_0 and k_0. Set $k^*(1) = k_0$ and $x^*(1) = x_0$.
10. Set time $t = 1$.
11. Find the ith and jth elements in the vectors of quantized k and x, which are closer to the values $k^*(t)$ and $x^*(t)$. If $(k(i), x(j)) = (k^*(t), x^*(t))$ then the state is a grid point, and $c^*(t)$ is read directly as $c(i, j, t)$. If $(k(i), x(j)) \neq (k^*(t), x^*(t))$, $c^*(t)$ is computer through bilinear interpolation using values of $c(i, j, t)$ at the closest grid points for k and x.
12. Check whether $t = T + 1$. If this equality is satisfied, go to point 15. Else, go to the next point.
13. Compute k^*_{t+1} and x^*_{t+1} according to the following equations:

$$k^*(t + 1) = (A + \delta - 1)k^*(t) - c^*(t) \tag{104}$$

$$x^*(t + 1) = Ak^*(t) - x^*(t)(\eta - \theta x^*(t) - 1) \tag{105}$$

14. Time t takes value $t + 1$. Go to point 11.
15. The value function is computed as follows: set time $t = T$ and an index $l = 1$. At final time T, the value function is computed according to the following formula:

$$V^*(T) = \frac{c^*(T)^{1-\sigma} - 1}{1-\sigma} - B\frac{x^*(T)\gamma}{\gamma} \tag{106}$$

16. Check whether $t = 0$. If so, end the program, otherwise go to the next point.
17. Time t takes values $T - l$.
18. The value function at time t is now

$$V^*(t) = \frac{c^*(t)^{1-\sigma} - 1}{1-\sigma} - B\frac{x^*(t)^{\gamma}}{\gamma} + \beta V^*(t+1) \tag{107}$$

19. The index l takes value $l + 1$. Go to point 16.

This computational procedure is very appealing for a number of reasons. First, because thorny questions about existence and uniqueness are avoided; as long as there is at least one feasible control sequence, then the direct-search procedure guarantees that the absolute maximum utility is achieved. Furthermore, extremely general types of systems equations and constraints can be handled. Constraints actually reduce the computational burden by decreasing the admissible sets of states and controls. Finally, the optimal control is obtained as a true feedback solution in which the optimal control for any admissible state and stage is determined. However, to the best of my knowledge there is not any algorithm able to identify whether multiple solutions exist, and this one makes no exception. Identify them may be difficult, because it requires repeated simulation of the same routine, and a bit of luck. If indeed multiple solutions exist, since in general they do not provide the same values of discounted flows of utility, the path which provides the highest value is generally chosen by the routine. So, most of the time, one does not even realise that multiple paths satisfying the first order conditions exist. They can only be found choosing appropriate grids, and this is a very difficult task because it requires first of all the knowledge about the existence of multiple solutions, and a good guess about the direction of the two paths. Finally, luck is always welcome.

Codes

In this section, I report the matlab codes that I used to draw the pictures and to compute the present value of the flow of utilities, in order to compare the two paths. The first program is the following, named

PROGRAM fp_main:

code:

```
clear

fp_parameters
fp_step1
fp_nextstates


fp_interpolation
save I1 I
save C1 C
```

PROGRAM fp_parameters:

code:

```
% parameters of the problem
B=10000000;
A=0.8;
rho=0.04;
theta=0.05;
eta=0.03;
sigma=3;
gamma=3;
delta=0.1;
beta=1-rho;

% definition of the grid for capital, pollution and consumption


kgrid=linspace(0.00,0.006,100); %This commnand creates a vector of 30
equidistant points from 0 to 0.1. The sintax is linspace(begin, end, number
of points)
xgrid=linspace(0.2,0.6,120);
cgrid=linspace(0.00,0.005,60);

% Number of periods
T=50;
```

PROGRAM fp_step1:

code:

```
% At time T (final), no control is applied because I assume that people
% consume what is left, so c(T)=(A-delta+1)k(T). So, for every state in
% kgrid and every x in xgrid, I compute and store C(i,j,t) (consumption
associated to the
% i-th state at time T) as well as I(i,j,t) (value function for each state
at
% time T).


% Initialisation of the matrix of solution for consumption:
C=NaN(length(kgrid),length(xgrid),T);

t=T;
for i=1:length(kgrid)
    for j=1:length(xgrid)
        I(i,j,t)=(((A-delta+1)*kgrid(i))^(1-sigma)-1)/(1-sigma)-B*xgrid(j)
^gamma/gamma;
        C(i,j,t)=(A-delta+1)*kgrid(i);
    end
end
```

PROGRAM fp_nextstates:

code:

```
for i=1:length(kgrid)
    for h=1:length(cgrid)
        K1(i,h)=(A-delta+1)*kgrid(i)-cgrid(h);
        if K1(i,h)<0 || K1(i,h)>max(kgrid)
            K1(i,h)=NaN;
        end
    end
end
```

```
for i=1:length(kgrid)
    for j=1:length(xgrid)
        X1(i,j)=A*kgrid(i)-xgrid(j)*(eta-theta*xgrid(j)-1);
        if X1(i,j)<0
               X1(i,j)=0;
        elseif X1(i,j)>max(xgrid)
                X1(i,j)=NaN;
        end
    end
end
```

PROGRAM fp_interpolation:

code:

```
for z=1:T-1
t=T-z;
for i=1:length(kgrid)
    for j=1:length(xgrid)
        for h=1:length(cgrid)

if isnan(K1(i,h))==0 && isnan(X1(i,j))==0

tmpk=abs(kgrid-K1(i,h));
[k k]=min(tmpk);
ck=kgrid(k);
tmpx=abs(xgrid-X1(i,j));
[x x]=min(tmpx);
cx=xgrid(x);

else
        I1(i,j,h,t)=NaN;
        V(i,j,h,t)=NaN;
end

    if ck==K1(i,h) && cx==X1(i,j)  % This means that X1 and K1 are grid
points
        I1(i,j,h,t)=I(k,x,t+1);
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif ck==K1(i,h) && X1(i,j)<xgrid(x) && x>1% This means that k1 is a
grid point but not x1. I have to interpolate between xgrid(x-1) and xgrid(x)
        I1(i,j,h,t)=I(k,x-1,t+1)+(X1(i,j)-xgrid(x-1))*(I(k,x,t+1)-I(k,x-1,
t+1))/(xgrid(x)-xgrid(x-1));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif ck==K1(i,h) && X1(i,j)>xgrid(x) && x<length(xgrid)% This means
that k1 is a grid point but not x1. I have to interpolate between xgrid(x)
and xgrid(x+1)
        I1(i,j,h,t)=I(k,x,t+1)+(X1(i,j)-xgrid(x))*(I(k,x+1,t+1)-I(k,x,t+1))/
(xgrid(x+1)-xgrid(x));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif cx==X1(i,j) && K1(i,h)<kgrid(k) && k>1% This means that X1 is a
grid point but not K1. I have to interpolate between kgrid(k-1) and kgrid(k)
        I1(i,j,h,t)=I(k-1,x,t+1)+(K1(i,h)-kgrid(k-1))*(I(k,x,t+1)-I(k-1,x,
t+1))/(kgrid(k)-kgrid(k-1));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif cx==X1(i,j) && K1(i,h)>kgrid(k) && k<length(kgrid) % This means
that X1 is a grid point but not K1. I have to interpolate between kgrid(k-1)
and kgrid(k)
        I1(i,j,h,t)=I(k,x,t+1)+(K1(i,h)-kgrid(k))*(I(k+1,x,t+1)-I(k,x,t+1))/
(kgrid(k+1)-kgrid(k));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
```

```
beta*I1(i,j,h,t);

    elseif K1(i,h)<kgrid(k) && X1(i,j)<xgrid(x) && k>1 && x>1
        % I need to interpolate between kgrid(k-1), kgrid(k), xgrid(x-1)
        % and xgrid(x)
        D=(xgrid(x)-xgrid(x-1))*(kgrid(k)-kgrid(k-1));
        I1(i,j,h,t)=I(k-1,x-1,t+1)/D*(xgrid(x)-X1(i,j))*(kgrid(k)-K1(i,
h))...
            +I(k-1,x,t+1)/D*(X1(i,j)-xgrid(x-1))*(kgrid(k)-K1(i,h))...
            +I(k,x-1,t+1)/D*(xgrid(x)-X1(i,j))*(K1(i,h)-kgrid(k-1))...
            +I(k,x,t+1)/D*(X1(i,j)-xgrid(x-1))*(K1(i,h)-kgrid(k-1));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif K1(i,h)>kgrid(k) && X1(i,j)<xgrid(x) && x>1 && k<length(kgrid)
         % I need to interpolate between kgrid(k), kgrid(k+1), xgrid(x-1)
         % and xgrid(x)
         D=(xgrid(x)-xgrid(x-1))*(kgrid(k+1)-kgrid(k));
         I1(i,j,h,t)=I(k,x-1,t+1)/D*(xgrid(x)-X1(i,j))*(kgrid(k+1)-K1(i,
h))...
             +I(k,x,t+1)/D*(X1(i,j)-xgrid(x-1))*(kgrid(k+1)-K1(i,h))...
             +I(k+1,x-1,t+1)/D*(xgrid(x)-X1(i,j))*(K1(i,h)-kgrid(k))...
             +I(k+1,x,t+1)/D*(X1(i,j)-xgrid(x-1))*(K1(i,h)-kgrid(k));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma
+ beta*I1(i,j,h,t);

    elseif K1(i,h)<kgrid(k) && X1(i,j)>xgrid(x) && k>1 && x<length(xgrid)
        % I need to interpolate between kgrid(k-1), kgrid(k), xgrid(x) and
        % xgrid(x+1)
        D=(xgrid(x+1)-xgrid(x))*(kgrid(k)-kgrid(k-1));
        I1(i,j,h,t)=I(k-1,x,t+1)/D*(xgrid(x+1)-X1(i,j))*(kgrid(k)-K1(i,
h))...
            +I(k-1,x+1,t+1)/D*(X1(i,j)-xgrid(x))*(kgrid(k)-K1(i,h))...
            +I(k,x,t+1)/D*(xgrid(x+1)-X1(i,j))*(K1(i,h)-kgrid(k-1))...
            +I(k,x+1,t+1)/D*(X1(i,j)-xgrid(x))*(K1(i,h)-kgrid(k-1));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    elseif K1(i,h)>kgrid(k) && X1(i,j)>xgrid(x) && k<length(kgrid) &&
x<length(xgrid)
        % I need to interpolate between kgrid(k), kgrid(k+1), xgrid(x), and
        % xgrid(x+1)
        D=(xgrid(x+1)-xgrid(x))*(kgrid(k+1)-kgrid(k));
        I1(i,j,h,t)=I(k,x,t+1)/D*(xgrid(x+1)-X1(i,j))*(kgrid(k+1)-K1(i,
h))...
            +I(k,x+1,t+1)/D*(X1(i,j)-xgrid(x))*(kgrid(k+1)-K1(i,h))...
            +I(k+1,x,t+1)/D*(xgrid(x+1)-X1(i,j))*(K1(i,h)-kgrid(k))...
            +I(k+1,x+1,t+1)/D*(X1(i,j)-xgrid(x))*(K1(i,h)-kgrid(k));
        V(i,j,h,t)=(cgrid(h)^(1-sigma)-1)/(1-sigma)-B*xgrid(j)^gamma/gamma +
beta*I1(i,j,h,t);

    end

end
        end
end


for i=1:length(kgrid)
    for j=1:length(xgrid)
        [c h]=max(V(i,j,:,t));
        C(i,j,t)=cgrid(h);
        I(i,j,t)=c;
    end
end
end
```

PROGRAM fp_submain:

code:

```
load C1.mat
load I1.mat
fp_parameters
fp_initialconditions
fp_retrievekx

fp_vstar

fp_plot
```

PROGRAM fp_initialconditions:

code:

```
%Declaration of the initial conditions. n represents the number of couples
%of initial conditions.

n=1;
k0=linspace(0.0025,0.0025,n);
x0=linspace(0.5,0.5,n);

IC=[k0;x0];
```

PROGRAM fp_retrievekx:

code:

```
% I generate a matrix of dimension 3,T. In the first row k*, in the second
% row x*, in the third row c*

sol=NaN(3,T,n);



z=1;
while z<=n
sol(1,1,z)=IC(1,z);
sol(2,1,z)=IC(2,z);
z=z+1;
end


 t=1;
 for z=1:n
 if sol(1,t,z)<=min(kgrid) || sol(1,t,z)>=max(kgrid) || sol(2,t,z)<=min
(xgrid) || sol(2,t,z)>=max(xgrid)
     display('Initial conditions out of the grid')
     display('Change the grid, or the initial conditions')
 end
 end
 z=1;
 while z<=n
     t=1;
    while t<=T-1
    if sol(1,t,z)>min(kgrid) && sol(1,t,z)<max(kgrid) && sol(2,t,z)>min
(xgrid) && sol(2,t,z)<max(xgrid)
```

```
tmpk=abs(kgrid-sol(1,t,z));
[k k]=min(tmpk);
ck=kgrid(k);
tmpx=abs(xgrid-sol(2,t,z));
[x x]=min(tmpx);
cx=xgrid(x);

else
        sol(1,t,z)=NaN;
        sol(2,t,z)=NaN;
        sol(3,t,z)=NaN;
end

if ck==sol(1,t,z) && cx==sol(2,t,z)
      % this means that x and k are grid points so c* is read directly
      sol(3,t,z)=C(k,x,t);
      sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
      if sol(1,t+1,z)<=mc/(A-delta)
          sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
      end
      sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);

elseif ck==sol(1,t,z) && sol(2,t,z)<xgrid(x)&& x>1
      % I need to interpolate between xgrid(x-1) and xgrid(x) - linear

          % interpolation since k is a grid point.
          sol(3,t,z)=C(k,x-1,t)+(sol(2,t,z)-xgrid(x-1))*(C(k,x,t)-C(k,x-1,
t))/(xgrid(x)-xgrid(x-1));
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);

    elseif ck==sol(1,t,z) && sol(2,t,z)>xgrid(x) && x<length(xgrid)
          % I need to interpolate between xgrid(x) and xgrid(x+1) - linear
          % interpolation since k is a grid point
          sol(3,t,z)=C(k,x,t)+(sol(2,t,z)-xgrid(x))*(C(k,x+1,t)-C(k,x,t))/
(xgrid(x+1)-xgrid(x));
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);

    elseif sol(1,t,z)<kgrid(k) && cx==sol(2,t,z) && k>1
          % I need to interpolate between kgrid(k-1) and kgrid(k) - linear
          % interpolation since x is a grid point
          sol(3,t,z)=C(k-1,x,t)+(sol(1,t,z)-kgrid(k-1))*(C(k,x,t)-C(k-1,x,
t))/(kgrid(k)-kgrid(k-1));
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);
```

```
    elseif sol(1,t,z)>kgrid(k) && cx==sol(2,t,z) && k<length(kgrid)
          % I need to interpolate between kgrid(k) and kgrid(k+1) - linear
          % interpolation since x is a grid point
          sol(3,t,z)=C(k,x,t)+(sol(1,t,z)-kgrid(k))*(C(k+1,x,t)-C(k,x,t))/...
(kgrid(k+1)-kgrid(k));
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);

    elseif sol(1,t,z)<kgrid(k) && sol(2,t,z)<xgrid(x) && k>1 && x>1
          % Bilinear interpolation between kgrid(k-1), kgrid(k), xgrid(x-1)
          % and xgrid(x)
          D=(kgrid(k)-kgrid(k-1))*(xgrid(x)-xgrid(x-1));
          sol(3,t,z)=C(k-1,x-1,t)*(kgrid(k)-sol(1,t,z))*(xgrid(x)-sol(2,t,...
z))/D...
              +C(k,x-1,t)*(sol(1,t,z)-kgrid(k-1))*(xgrid(x)-sol(2,t,z))/D...
              +C(k-1,x,t)*(kgrid(k)-sol(1,t,z))*(sol(2,t,z)-xgrid(x-1))/D...
              +C(k,x,t)*(sol(1,t,z)-kgrid(k-1))*(sol(2,t,z)-xgrid(x-1))/D;
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);


          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);


    elseif sol(1,t,z)<kgrid(k) && sol(2,t,z)>xgrid(x) && k>1 && x<length...
(xgrid)
          % Bilinear interpolation between kgrid(k-1), kgrid(k), xgrid(x),
          % xgrid(x+1)
          D=(kgrid(k)-kgrid(k-1))*(xgrid(x+1)-xgrid(x));
          sol(3,t,z)=C(k-1,x,t)*(kgrid(k)-sol(1,t,z))*(xgrid(x+1)-sol(2,t,...
z))/D...
              +C(k,x,t)*(sol(1,t,z)-kgrid(k-1))*(xgrid(x+1)-sol(2,t,z))/D...
              +C(k-1,x+1,t)*(kgrid(k)-sol(1,t,z))*(sol(2,t,z)-xgrid(x))/D...
              +C(k,x+1,t)*(sol(1,t,z)-kgrid(k-1))*(sol(2,t,z)-xgrid(x))/D;
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);

    elseif sol(1,t,z)>kgrid(k) && sol(2,t,z)<xgrid(x) && x>1 && k<length...
(kgrid)
          % Bilinear interpolation between kstar(k), kstar(k+1), xstar(x-1)
          % and xstar(x)
          D=(kgrid(k+1)-kgrid(k))*(xgrid(x)-xgrid(x-1));
          sol(3,t,z)=C(k,x-1,t)*(kgrid(k+1)-sol(1,t,z))*(xgrid(x)-sol(2,t,...
z))/D...
              +C(k+1,x-1,t)*(sol(1,t,z)-kgrid(k))*(xgrid(x)-sol(2,t,z))/D...
              +C(k,x,t)*(kgrid(k+1)-sol(1,t,z))*(sol(2,t,z)-xgrid(x-1))/D...
              +C(k+1,x,t)*(sol(1,t,z)-kgrid(k))*(sol(2,t,z)-xgrid(x-1))/D;
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);
```

```
    elseif sol(1,t,z)>kgrid(k) && sol(2,t,z)>xgrid(x) && k<length(kgrid) &&
x<length(xgrid)
          D=(kgrid(k+1)-kgrid(k))*(xgrid(x+1)-xgrid(x))
          sol(3,t,z)=C(k,x,t)*(kgrid(k+1)-sol(1,t,z))*(xgrid(x+1)-sol(2,t,
z))/D...
              +C(k+1,x,t)*(sol(1,t,z)-kgrid(k))*(xgrid(x+1)-sol(2,t,z))/D...
              +C(k,x+1,t)*(kgrid(k+1)-sol(1,t,z))*(sol(2,t,z)-xgrid(x))/D...
              +C(k+1,x+1,t)*(sol(1,t,z)-kgrid(k))*(sol(2,t,z)-xgrid(x))/D;
          sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          if sol(1,t+1,z)<=mc/(A-delta)
              sol(3,t,z)=(A-delta+1)*sol(1,t,z)-(mc/(A-delta));
              sol(1,t+1,z)=(A-delta+1)*sol(1,t,z)-sol(3,t,z);
          end
          sol(2,t+1,z)=A*sol(1,t,z)-sol(2,t,z)*(eta-theta*sol(2,t,z)-1);


  end
 t=t+1;
end
sol(3,T,z)=sol(1,T,z);
z=z+1;
end
```

PROGRAM fp_vstar:

code:

```
% This program computes - backward in time - the value functions for the
% optimal path found in the program l_retreivekx4 given the initial
% conditions IC.
z=1;
while z<=n
t=T;

Vstar(T,z)=(sol(1,T,z)^(1-sigma)-1)/(1-sigma)-(B*sol(2,T,z)^gamma)\gamma;
for i=1:T-1
    Vstar(t-i,z)=(sol(3,t-i,z)^(1-sigma)-1)/(1-sigma)-(B*sol(2,t-i,z)^gamma)
/gamma + beta*Vstar(t-i+1,z);

end
z=z+1;
end
```

PROGRAM fp_plot:

code:

```
%plotting tools

l_parameters5
domain=0:0.001:0.6;
C=1/(A-delta)*(A*B/(A-delta-rho))^(-1/sigma);

for j=1:length(domain)
F1(j)=domain(j)*(eta-theta*domain(j))/A;
F2(j)=C*domain(j)^((1-gamma)/sigma)*(eta-2*theta*domain(j)+rho)^(1/sigma);
end

plot(domain,F1, ':',domain,F2,':')
axis([0 0.6 0 0.007])
hold on
```

```
% Creation of the variables for plotting (cancel the last three periods
% because of not beautiful graph

solplot=sol;

for z=1:n
plot(solplot(2,:,z),solplot(1,:,z))
end
xlabel('Pollution')
ylabel('Capital')
hold on
```

References

1. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
2. Bellman, R., Dreyfus, S.: Applied Dynamic Programming. Princeton University Press, Princeton (1962)
3. Caputo, M.: Foundations of Dynamic Economic Analysis. Optimal Control Theory and Applications. Cambridge University Press, Cambridge (2005)
4. Dasgupta, S.: Control of Resources. Basil Blackwell, Oxford (1982)
5. Dasgupta, S., Laplante, B., Wang, H., Wheeler, D.: Confronting the environmental Kuznets curve. J. Econ. Perspect. **16**(1), 147–168 (2002)
6. Deissenberg, Ch., Feichtinger, G., Semmler, W., Wirl, F.: Multiple equilibria, history dependence, and global dynamics in intertemporal optimization models. In: Barnett, W., Deissenberg, Ch., Feichtinger, G. (eds.) Economic Complexity: Non-linear Dynamics, Multi-agents Economies, and Learning, ISETE, vol. 14, pp. 91–122. Elsevier, Amsterdam (2003)
7. Forster, B.: Optimal pollution control with a nonconstant exponential rate of decay. J. Environ. Econ. Manag. **2**, 1–6 (1975)
8. Holling, C.S.: Resilience and stability of ecological systems. Rev. Ecol. Syst. **4**, 1–23 (1973)
9. Keeler, E., Spence, M., Zeckhauser, R.: The optimal control of pollution. J. Econ. Theory **4**, 19–34 (1972)
10. Larson, R.: A survey of dynamic programming computational procedures. In: IEEE Transaction on Automatic Control, pp. 767–774 (1967)
11. Nordhaus, W.: A review of the Stern review on the economics of climate change. J. Econ. Lit. **45**, 686–702 (2007)
12. Ryder, H.E., Heal, G.M.: Optimal growth with intertemporally dependent preferences. Rev. Econ. Stud. **40**(1), 1–31 (1973)
13. Seierstad, A., Sydsæter, K.: Optimal Control Theory with Economic Applications. North Holland, Amsterdam (1987)
14. Stern, N.: The Economics of Climate Change. The Stern Review. Cambridge University Press, Cambridge and New York (2007)
15. Stokey, N.: Are there limits to growth? Int. Econ. Rev. **39**(1), 1–31 (1998)
16. Tahvonen, O., Withagen, C.: Optimality of irreversible pollution accumulation. J. Econ. Dyn. Control **20**, 1775–1795 (1996)

Stochastic Modelling of Biochemical Networks and Inference of Model Parameters

Vilda Purutçuoğlu

Abstract There are many approaches to model the biochemical systems deterministically or stochastically. In deterministic approaches, we aim to describe the steady-state behaviours of the system, whereas, under stochastic models, we present the random nature of the system, for instance, during transcription or translation processes. Here, we represent the stochastic modelling approaches of biological networks and explain in details the inference of the model parameters within the Bayesian framework.

Keywords Stochastic modelling · Bayesian inference · Diffusion bridge method Particle filtering method

1 Introduction

A list of reactions which describes the biological process under different constrains can be modelled by distinct techniques [3, 17]. The Boolean, differential equation and the stochastic modelling are the major approaches. The *Boolean approach* denotes the activation of the system via on, i.e., fully expressed, or off, i.e., not fully expressed, positions. For a given state, the system passes to a next state deterministically assuming that all genes simultaneously change their states. Since the state number is taken as finite, it is useful to get a first impression for complex systems in a large state space [3]. In the *differential equation models*, the states are indicated as continuous concentrations changing by time according to nonlinear ordinary differential equations (ODEs) and rate of reactions. In order to estimate the model parameters which are the rate of reactions, two main approaches are suggested. The first method is based on solving the set of differential equations simultaneously by equating each of them to zero under the assumption that one state is converted to another state

V. Purutçuoğlu (✉)
Department of Statistics, Middle East Technical University,
06800 Ankara, Turkey
e-mail: vpurutcu@metu.edu.tr

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_18

without observing any net change. So the states are taken as equilibrium by setting all derivatives to zero. As the underlying function is typically nonlinear, the unique solution of the system is not solvable. Hence, the second approach assumes that these derivatives, i.e., set of differential equations, can be approximated via different methods such linear models, neural networks or recurrent artificial neural networks models [3]. Finally in the *stochastic models*, the exact number of molecules is used to describe the states of the systems, where the present number of molecules of each type causing the system moves to the next state probabilistically. Hence, the reaction type and its time are defined probabilistically.

In this chapter, we explain the stochastic models and the inference of model parameters in details. The very general idea of stochastic models is to represent the random feature of the biological processes by means of the number of molecules of the species. Since these models apply very detailed information about the systems, the collected knowledge becomes more comprehensive than the outputs of the models based on the ordinary differential equations, accordingly, boolean models. In stochastic model, each biological event is shown via a chemical equation as below.

$$u_1 R_1 + u_2 R_2 + \ldots + u_{nr} R_r \xrightarrow{c} v_1 P_1 + v_2 P_2 + \ldots + v_{np} P_p, \tag{1}$$

where u_i $(i = 1, \ldots, r)$ and v_j $(j = 1, \ldots, p)$ stand for the *stoichiometric coefficient* of the reactant R and the *stoichiometric coefficient* of the product P, respectively. The stoichiometric coefficient indicates the necessary number of molecules either consumed (if it is on the left-hand side of the equation) or produced (if it is on the right-hand side of the equation) in a single reaction step. Therefore, the chemical meaning of Eq. 1 is that u_i amount of R_i molecules collides with each other and produces v_j amount of P_j molecules when the molecules move by the Brownian motion [3]. As a result, at the end of each reaction, there is a change in the system with the amount of $s = v - u$ that is also called as the *net effect*. Finally, c shows the *stochastic reaction rate constant* which implies the speed of the reaction and the model parameter in modelling of this reaction via different approaches.

Accordingly, if we represent the stoichiometric coefficient of each reaction as a vector, resulting in a matrix for the set of reactions, also known as the *reaction list*, the net effect matrix S can be denoted via $S = V - U$ where V and U are the product and the reactant matrix, respectively, the reaction rate constants can be denoted by a vector Θ whose entries display the reaction rate constant for each reaction.

In order to model such reactions as in Eq. 1, we can choose different models. In this chapter, we merely focus on the stochastic modelling approaches and alternative methods for the inference of model parameters.

2 Stochastic Modelling Approaches

There are three main methods to stochastically model the biochemical system. The first method is called as the *Langevin model* where a differential equation in concentrations is extended by adding a stochastic noise term. The second method defines a model from a deterministic differential equation for the dynamics of a probability distribution that are known as the *Fokker–Plank equation* and also the *diffusion approximation model*. In these two models, the former is obtained from the extension of the differential equation and the latter is derived from the simplification of the fully stochastic model so that for a given state, the derived differential equations of the joint density for the number of molecules and time are only the functions of the change in time [3]. Lastly, the third method is named as the *inhomogeneous Poisson process model*, that is generated from the Gillespie algorithm [9, 12] in the calculation of the likelihood equations [26]. Below, we represent each of the alternative models in details.

2.1 Langevin Model

The Langevin model can be considered as the extension of the differential equations approach which includes a noise term as

$$\frac{d}{dt}Y(t) = \mu(Y, \Theta) + W(t), \tag{2}$$

where t stands for the time and $W(t)$ indicates the time-dependent stochastic process defined by the Brownian motion over time. Thereby, $\mu(Y, \Theta)$ denotes the mean changes in states as a function of the state Y and model parameters Θ. In Eq. 2, the calculations of the number of molecules are tractable for linear μ. But, if this deterministic part is non-linear, it can be inferred by linear approximation techniques [25] whose solutions are not unique. Furthermore, the model in Eq. 2 can be extended by adding a noise term dependent on state into the last term, $W(t)$. Under such extended model, the equation is solved via the Itô or Stratonovich integrals [21, 25].

2.2 Diffusion Approximation Model

In a gene network model, under the assumption of continuous number of molecules Y at continuous time t, the probability distribution of states and time, i.e., P(number of molecules, t), can be explained by differential equations models. Here, the Fokker–Planck equation, also known as the *Smoluchowki equation*, the *second Kolmogorov equation*, or the *generalized diffusion equation*, converts the stochastic expression to

the differential equation by assuming a continuous density on the number of molecules through time whose derivation is based on the following master equation [10, 11, 24].

$$\frac{\partial P(Y,t)}{\partial t} = \sum_{j=1}^{r} \{h_j(Y - s_j; \Theta) P(Y - s_j, t) - h_j(Y; \Theta) P(Y, t)\} \tag{3}$$

in which the n-dimensional vector $Y = (Y_1, Y_2, \ldots, Y_n)$ indicates the state of the system at time t and r-dimensional vector $\Theta = (c_1, c_2, \ldots, c_r)$ show the stochastic reaction rate constants. Hereby, $P(Y, t)$ in Eq. 3 denotes the probability distribution of states which is described by discrete number of molecules and continuous time t. Finally, S and h_j describe the net effect matrix as used previously, and the hazard of the jth reaction for the total number of r reactions and n substrates, in order. The hazard, also known as the *rate law of reaction*, represents the product of the number of distinct molecular reactant combinations available in the state Y for reaction j with stochastic rate constant Θ. For instance, the hazard of the reaction given in Eq. 3 at time t can be computed by $h(Y, \Theta) = c \times \binom{[R_1]}{u_1} \times \ldots \times \binom{[R_r]}{u_r}$ where $\binom{[R_i]}{u_i}$ describes the molecular combination while $[R_i]$ shows the number of present molecules of the ith substrates and u_i displays its associated stoichiometric coefficient as stated previously.

Accordingly, $h_j(Y - s_j; \Theta) P(Y - s_j, t)$ in Eq. 3 implies the probability of the jth reaction over the time interval $[t, t + dt]$ so that the system changes the states from $Y - s_j$ to Y [24, 25] and the term $h_j(Y, \Theta) P(Y, t)$ denotes the probability of no reaction in the same time interval so that the system is in the same state. Then, if Eq. 3 is defined in terms of jump sizes and the expression is approximated via the second order Taylor series expansion and written in terms of jump moments, the following equation, also called as the *Fokker–Planck equation*, is obtained.

$$\frac{\partial P(Y,t)}{\partial t} = -\frac{\partial}{\partial Y}\left(\int \lambda h(Y;\Theta) d\lambda\right) P(Y,t) + \frac{1}{2}\frac{\partial^2}{\partial Y^2}\left(\int \lambda^2 h(Y;\Theta) d\lambda\right) P(Y,t) \tag{4}$$

where λ is the jump size when the system has the starting point Y', i.e., $\lambda = Y - Y'$. Furthermore, in this equation, the first part on the right-hand side is known as the *transport*, the *convection*, or the *drift* term and the second term on the same side is named as the *diffusion* or the *fluctuation* term.

On the other hand, if Eq. 4 is expressed under the infinitesimal means $\mu_i(Y, \Theta)$ and the second moments of the jump of states $\beta_{ij}(Y, \Theta)$ for $i, j = 1, \ldots, n$ as the index of substrates during the time change τ, Fokker–Planck equation can be defined as

$$\frac{\partial}{\partial t}P(Y,t) = -\sum_{i=1}^{n}\frac{\partial}{\partial Y}\{\mu_i(Y,\Theta)P(Y,t)\} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial Y_i \partial Y_j}\{\beta_{ij}(Y,\Theta)P(Y,t)\} \tag{5}$$

where

$$\mu_i(Y,\Theta) = \lim_{\tau\to 0}\frac{1}{\tau}E[\{Y_i(t+\tau) - Y_i(t)\}|Y(t) = Y],$$

and

$$\beta_{ij}(Y,\Theta) = \lim_{\tau\to 0}\frac{1}{\tau}\mathrm{Cov}[\{Y_i(t+\tau) - Y_i(t)\},\{Y_j(t+\tau) - Y_j(t)\}|Y(t) = Y]$$

by using the expectation $E(.)$ and the covariance $Cov(.)$ expressions. On the other hand, if Eq. 5 is presented by the Itô diffusion, then we can obtain the *diffusion approximation model* of the system via

$$dY(t) = \mu(Y,\Theta)dt + \beta^{\frac{1}{2}}(Y,\Theta)dW(t) \tag{6}$$

in which $\mu(Y,\Theta) = S'h(Y,\Theta)$ and $\beta(Y,\Theta) = S'\mathrm{diag}\{h(Y,\Theta)\}S$ show the *mean*, or *drift*, and *variance*, or *diffusion* matrices, respectively. Similar to previous expressions, Y and $\Theta = (c_1, c_2, \ldots, c_r)'$ denote the state and the parameter vector. $dW(t)$ is the change of a Brownian motion over time and S means the net effect matrix. Finally, we set λ and τ to $\lambda(Y) = dY$ and $\tau = dt$, in order [13].

However, in practice, a finitely observed sample path as defined in the diffusion approximation in Eq. 6, is intractable because of the missing states between observed end points. Therefore, we need to implement the discretized version of the diffusion model under the assumption of discrete jumps at a large set of discrete time points. This discretized version of the diffusion approximation, shown in Eq. 7, is called the *Euler–Maruyama approximation*.

$$\Delta Y_t = \mu(Y_t,\Theta)\Delta t + \beta^{\frac{1}{2}}(Y_t,\Theta)\Delta W_t \tag{7}$$

where ΔW_t implies an n-dimensional independent identically distributed Brownian random vector generated by normal distribution with mean zero and covariance-variance as the product of the identity matrix I and discrete time interval Δt, i.e., $\Delta W_t \sim N(0, I\Delta t)$. $Y = (Y_1, \ldots, Y_n)$ presents the state of the system at time t and $\Theta = (c_1, \ldots, c_r)'$ stands for the parameter vector while n and r are the total number of substrates and the total number of reactions in the system, respectively, as used beforehand [26]. The Euler–Maruyama approximation enables us to derive a complete-data likelihood for the sample path.

2.3 Inhomogeneous Poisson Process Model

Under stochastic models, it is assumed that the sample path of each gene is observed through time similar to the idea of the Gillespie algorithm which can exactly simulate the biological network stochastically [9]. Hereby, the inhomogeneous poisson process is based on the application of the Gillespie algorithm that is derived from the chemical master equation [9, 25]. Thus, if we perform the same sampling plan of the Gillespie algorithm in a stochastic model, the complete data likelihood $L(\Theta, y)$ can be expressed as below.

$$L(\Theta, Y) = \left[\prod_{i=1}^{r} h_i(y_{i-1}(t), c_i)\right] \exp\left\{-\int_0^T h_0(y(t), \Theta)dt\right\} \tag{8}$$

for a model between the time interval $[0, T]$. In Eq. 8, $Y = (y_1, \ldots, y_n)$ and $\Theta = (c_1 \ldots, r)$ present the state and the reaction rate vector, respectively. Additionally, h_i and dt indicate the hazard and the change in time, in order. As a result, under the assumption of the gamma distribution as the prior of reaction rates, we can derive the posterior of the jth reaction rate ($j = 1, \ldots, r$) as the following expression.

$$c_j|y \sim \Gamma\left[a_j + r_j, b_j + \int_0^T g_j(y(t))dt\right] \tag{9}$$

in which a_j and b_j denote the given parameters of the gamma distribution $\Gamma[.,.]$ and r_j describes the number of the jth ($j = 1, \ldots, r$) reaction in the sample path Y. Furthermore, $g_j(y(t)) = h_j(y(t), c_j)/c_j$ [26]. Hence, Eq. 9 shows that for the given state, the reaction rate constants have a known distribution, resulting in inference of these model parameters via the Gibbs sampling. But, such calculation can be feasible if we get the complete sample path. However, in practice, we can merely observe it partially. Therefore, an approximate model which is based on the computation of the end points of the correct sample path obtained from the Gillespie algorithm in its likelihood expression can be applicable. Hereby, the underlying approximate modelling is called as the *inhomogeneous Poisson process model* that can be also described as the nonlinearly re-scaled homogeneous Poisson process by time [26].

This approach accepts that the states, which are the functions of hazards, are known at the end points $y(0)$ and $y(1)$, and a possible true path can be constructed by approximately simulating hazards from the independent inhomogeneous Poisson processes with rates linear between the rates at the end points by

$$N \sim Poi(\lambda) \text{ and } Y|N \sim B(N, p), \text{ then } Y \sim Poi(\lambda p).$$

In this sampling plan, λ and p denote the constant rate of the Poisson process and the reaction probability in a small time interval, respectively. Additionally, N shows the number of reactions and Y is the true sample path within this time [26]. Finally, $B(.,.)$ implies the binomial distribution with the given parameters.

As a result, the complete-data likelihood under the inhomogeneous Poisson process model can be described as

$$L_A(\Theta, y) = \left[\prod_{j=1}^{r} \lambda_j(t)\right] \exp\left\{-\frac{1}{2}[h_0(y_0(t), \Theta) + h_0(y_1(t), \Theta)]\right\} \tag{10}$$

in which $\lambda_j(t) = (1-t)h_j(y_0(t), \Theta) + th_j(y_1(t), \Theta)$ for $j = 1, \ldots, r$ and means the Poisson rate of each jth reaction whose reaction rate is equal to c_j under $\Theta = (c_1, \ldots, c_r)$ for $j = 1, \ldots, r$. Thus, if we are interested in inference of the model parameters $\Theta = (c_1, \ldots, c_r)$ based on Eqs. 8 and 10, we can construct a Bayesian sampling plan [26]. But since in the current literature, the estimation of $\Theta = (c_1, \ldots, c_r)$ via the inhomogeneous Poisson process model is performed for small systems, we will explain the calculation of the inference methods based on the diffusion approximation, and accordingly, based on the Langevin model. Because, these models can be applicable for both small and realistically large systems.

3 Inference of Model Parameters Based on Diffusion Approximation Model

The major challenge in stochastic modelling is the application of the model in large systems. There are three main approaches in order to estimate the model parameters based on both the diffusion approximation and the Langevin models. As explained previously, these two models indicate the same mathematical expressions and therefore, the inference of their model parameters is conducted via the same approaches. The *Metropolis-within-Gibbs algorithm by columnwise update*, the *modified diffusion bridge method* and the *particle filtering technique*, that are based on Bayesian computations, are the underlying three main methods.

In the estimation of reaction rates in a biochemical system via these methods, we use the following state matrix Y which consists of the observations at the given time points $t_i = t_0, \ldots, t_T$. But for a realistic inference, [13] consider the partially observed states, rather than fully observed Y. Hence, the matrix below is composed of both unobserved Z and observed X parts so that $Y = (X, Z)$. Here, X indicates a $(n_x \times t_T)$-dimensional observed sub-matrix with n_x amount of observed species and Z presents a $(n_z \times t_T)$-dimensional unobserved sub-matrix with n_z amount of unobserved species. Accordingly, $n = n_x + n_z$ shows the total number of substrates in the system.

$$Y = \begin{bmatrix} X_1(t_0) & X_1(t_1) & X_1(t_2) & \dots & X_1(t_T) \\ X_2(t_0) & X_2(t_1) & X_2(t_2) & \dots & X_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n_x}(t_0) & X_{n_x}(t_1) & X_{n_x}(t_2) & \dots & X_{n_x}(t_T) \\ Z_1(t_0) & Z_1(t_1) & Z_1(t_2) & \dots & Z_1(t_T) \\ Z_2(t_0) & Z_2(t_1) & Z_2(t_2) & \dots & Z_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{n_z}(t_0) & Z_{n_z}(t_1) & Z_{n_z}(t_2) & \dots & Z_{n_z}(t_T) \end{bmatrix}.$$

On the other hand, since in practice, the diffusion approximation is performed via its discretized version, called the Euler–Maruyama approximation, as explained in Sect. 2.2, the time interval between the subsequent two observed time points needs to be very small. But, as the data at hand are typically very limited, [6] proposes the data augmentation technique for states Y in order to decrease the large biased due to the discretization. This strategy is already used for univariate diffusion model within an importance sampling [5] and a simulation-based approach [13, 18] implements it for the inference of the multivariate diffusion model.

Thus, the following extended matrix of Y is generated by adding *latent* or *augmented* states between each pair of observed time points of original Y.

$$Y = \begin{bmatrix} x_1(t_0) & X_1(t_1) & \dots & X_1(t_{m-1}) & x_1(t_m) & X_1(t_{m+1}) & \dots & X_1(t_{mT-1}) & x_1(t_T) \\ x_2(t_0) & X_2(t_1) & \dots & X_2(t_{m-1}) & x_2(t_m) & X_2(t_{m+1}) & \dots & X_2(t_{mT-1}) & x_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{d_1}(t_0) & X_{d_1}(t_1) & \dots & X_{d_1}(t_{m-1}) & x_{d_1}(t_m) & X_{d_1}(t_{m+1}) & \dots & X_{d_1}(t_{mT-1}) & x_{d_1}(t_T) \\ Z_1(t_0) & Z_1(t_1) & \dots & Z_1(t_{m-1}) & Z_1(t_m) & Z_1(t_{m+1}) & \dots & Z_1(t_{mT-1}) & Z_1(t_T) \\ Z_2(t_0) & Z_2(t_1) & \dots & Z_2(t_{m-1}) & Z_2(t_m) & Z_2(t_{m+1}) & \dots & Z_2(t_{mT-1}) & Z_2(t_T) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{d_2}(t_0) & Z_{d_2}(t_1) & \dots & Z_{d_2}(t_{m-1}) & Z_{d_2}(t_m) & Z_{d_2}(t_{m+1}) & \dots & Z_{d_2}(t_{mT-1}) & Z_{d_2}(t_T) \end{bmatrix},$$

where x_i and X_i display the observed data and the augmented data, respectively, by observed components. Thereby, each $Y_i = (X_i, Z_i)$ denotes the ith column of Y assuming that the number of augmented states m is the same in every pair of observed time points.

Hence, the aim of all inference approaches is based on both estimating these missing states and reaction rate constants, which are the main parameters of interest in systems, via different schemes. Below, we present the details of the update of Y and Θ, in details.

3.1 Metropolis-Within-Gibbs Algorithm by Columnwise Updates

In the update of the state matrix Y via this algorithm, the conditional posterior of each state is described via the multivariate normal distribution. But, the candidate generator of each state is simulated from distinct multivariate normal densities based on whether the underlying state is partially observed as the first and the last columns, i.e., Y_0 and Y_T, as well as for the observed time points or fully unobserved as certain middle columns such as Y_{m-1} or Y_{m+1}. Accordingly the update of the states is computed via the Metropolis-Hastings algorithm for all columns apart from the last one and the last column is generated via the Gibbs sampling [7, 8].

For instance, if the next column i is partially observed but not equal to neither 0 nor T, we need to merely sample a candidate value for Z_i conditional on $X_i = x_i$. Thus the candidate state Y^* is described as below.

$$Y_i^* = \begin{pmatrix} X_i^* \\ Z_i^* \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{1}{2}(X_{i-1} + X_{i+1}) \\ \frac{1}{2}(Z_{i-1} + Z_{i+1}) \end{pmatrix}, \begin{pmatrix} \beta_{i-1}^{xx} & \beta_{i-1}^{xz} \\ \beta_{i-1}^{zx} & \beta_{i-1}^{zz} \end{pmatrix} \frac{1}{2}\Delta t\right). \tag{11}$$

In this matrix Z_i^* conditional on $X_i = x_i$ is given as

$$\eta_{Z_i^*} = \frac{1}{2}(Z_{i-1} + Z_{i+1}) + \beta_{i-1}^{zx}(\beta_{i-1}^{xx})^{-1}(x_i - \frac{1}{2}[X_{i-1} + X_{i+1}]) \tag{12}$$

$$\Sigma_{Z_i^*} = \frac{1}{2}\Delta t(\beta_{i-1}^{zz} - \beta_{i-1}^{zx}(\beta_{i-1}^{xx})^{-1}\beta_{i-1}^{xz}), \tag{13}$$

where $\eta_{Z_i^*}$ and $\Sigma_{Z_i^*}$ are the associated mean and covariance-variance matrix, respectively. Here, $\beta_{i-1}^{zx} = \beta(Y_{i-1}^{zx}, \Theta)$, $\beta_{i-1}^{xx} = \beta(Y_{i-1}^{xx}, \Theta)$ has full rank, $\beta_{i-1}^{zz} = \beta(Y_{i-1}^{zz}, \Theta)$, and $\beta_{i-1}^{xz} = \beta(Y_{i-1}^{xz}, \Theta)$. Then, we decide on accepting or rejecting step by computing the probability as follows.

$$\alpha(Z_i^*|Z_i) = \min\left\{1, \frac{p(Z_i^*|x_i, Y_{i-1}, Y_{i+1}, \Theta)q(Z_i|x_i, Y_{i-1}, Y_{i+1}, \Theta)}{p(Z_i|x_i, Y_{i-1}, Y_{i+1}, \Theta)q(Z_i^*|x_i, Y_{i-1}, Y_{i+1}, \Theta)}\right\}. \tag{14}$$

for the transition kernels $q(.)$ under the current and proposal states.

On the other hand, if the state is a latent state, then the candidate generator Y^* can be simulated from Eq. 15 as follows.

$$Y^* \sim N\left(\frac{1}{2}(Y_{i-1} + Y_{i+1}), \frac{1}{2}\Delta t\beta(Y_{i-1}, \Theta)\right) \tag{15}$$

which implies $q(.|Y_{i-1}, Y_{i+1}, \Theta)$ and accordingly converges pointwise to $\pi(.|Y_{i-1}, Y_{i+1}, \Theta)$ as long as Δt goes to 0 as stated beforehand. The algorithm updates the columns regarding to the probability

$$\alpha(Y_i^*|Y_i) = \min\left\{1, \frac{p(Y_i^*|Y_{i-1}, Y_{i+1}, \Theta)q(Y_i|Y_{i-1}, Y_{i+1}, \Theta)}{p(Y_i|Y_{i-1}, Y_{i+1}, \Theta)q(Y_i^*|Y_{i-1}, Y_{i+1}, \Theta)}\right\}. \tag{16}$$

similar to the acceptance probability given in Eq. 14.

Finally to update Θ, the normal distribution can be chosen due to the simplicity and the random walk algorithm can be performed for the inference of Θ. Hereby, the conditional posterior is calculated by

$$\pi(\Theta|Y) \propto \prod_{i=1}^{T} f(Y_i|Y_{i-1}, \Theta)\pi(\Theta) \tag{17}$$

in which $\pi(\Theta)$ implies to the prior distribution of the reaction rates. If there is no biological knowledge about the possible values of these parameters apart from their positivity condition, the prior can be uninformative, such as uniform, Jeffrey prior [7] or exponential with small rate like 1 [19]. In our derivation, we prefer Exp(1) to guarantee the positivity constraints of Θ. Accordingly, a candidate Θ^* is accepted with respect to the given acceptance probability below.

$$\alpha(\Theta, \Theta^*|Y) = \min\left\{1, \frac{\pi(\Theta^*|Y)}{\pi(\Theta|Y)}\right\}, \tag{18}$$

where Θ^* is produced from $\Theta^* = \Theta + w$ when w presents the r-dimensional vector of perturbations for Θ such that $w = (w_1, \ldots, w_j, \ldots, w_r)$ $(j = 1, \ldots, r)$. In the estimation, each w_j can be generated from $w_j \sim N(0, \gamma_j)$ in which γ_j indicates the variance of w_j for each reaction and implicitly controls the mixing property of the MCMC algorithm. Therefore, γ_j is also called as the *tuning parameter* of Θ. For good mixing in random walk chains under the univariate case, an acceptance rate of around 24% is optimal [7, 23]. But for large number of parameters, this value can be decreased [19, 20].

In the update of Θ, the acceptance probability in Eq. 18 is compared with a standard uniform random value u. If $u < \alpha(\Theta, \Theta^*|Y)$, the candidate reaction rates are replaced by the current rates, otherwise, Θ^* is rejected in that iteration. This updating scheme continues until the convergence is achieved for all reaction rates.

On conclusion, even though the Metropolis-within-Gibbs algorithm is successful in inference of the reaction rates under large missing values, the number of latent states added between each pair of observed states in order to decrease the bias in the discretized version of the diffusion process may cause high correlation. To solve the underlying problem, different sampling plans can be applied as explained in the following parts or a measurement error can be included in our diffusion model [16].

3.2 *Modified Diffusion Bridge Method*

In order to unravel the challenge of high correlation between states, the sampling plan can be extended by constructing a diffusion bridge between observed time points [14, 15]. Briefly, the scheme of the diffusion bridge is still based on the Metropolis-within-Gibbs algorithm in the updates of states. But different from the columnwise method, it updates $(2m + 1)$ states simultaneously where m is the number of latent states between each pair of observed time points and every mth column shows a partially observed column in the state matrix Y. With more details, in this estimation method, the jth column of Y, Y_j, is updated by using M and M^+ terms. Here, $M = j + m$ and $M^+ = M + m$. Furthermore, since each column of Y is described by $Y = (x, Z)$ when x and Z denote the observed and unobserved components, the conditional posterior density of the sample path between j and M^+ columns given the observed time points ($t \neq 0$ and $t \neq T$) and reaction rates Θ can be written as

$$\pi(Y_{j+1}, \ldots, Z_M, \ldots, Y_{M^+-1} | Y_j, x_M, Y_{M^+}, \Theta) \propto \prod_{i=j}^{M^+-1} \pi(Y_{i+1} | Y_i, \Theta) \tag{19}$$

in which $Y_M = (x_M, Z_M)$ and $j = 0, m, \ldots, T - 2m$. T is the total time points after adding the latent or augmented states. For instance, for the initial state, this density can be simulated by taking the next $(m - 1)$ columns via

$$\pi(Z_0, Y_1, Y_{m-1} | x_0, Y_m, \Theta) \propto \prod_{i=0}^{m-1} \pi(Y_{i+1} | Y, \Theta)$$

for the unobserved part Z_0 in which $Y_0 = (x_0, Z_0)$. Then, by including the conditional density of the first and last columns within a discretized diffusion model, the states can be generated from a multivariate normal distribution in two different ways. Here, we present one of these methods. Because the second method is the extension of the first approach by also adding Y^{M^+} states in the calculation of this one and more details are reported in Sect. 3.3 during the updating scheme of the states and reaction rates in Eq. 24.

Accordingly, in the first approach, each Y_{k+1} column is inferred from Y_k, x_M, and Θ where $k = j, \ldots, M - 2$. Under this constraint, the density of Y_{k+1} can be obtained by

$$\pi(Y_{k+1} | Y_k, x_M, \Theta) \sim N(Y_k + \mu_k^*, \Sigma_k^*) \tag{20}$$

where

$$\mu_k^* = \begin{pmatrix} (x_M - X_k)/(M-k) \\ \left\{\mu(Z_k, \Theta)(M-k)\Delta t + \beta(Y_k^{zx}, \Theta)\beta(X_k, \Theta)^{-1}(x_M - [X_k + \mu(X_k, \Theta)(M-k-1)\Delta t])\right\}/(M-k) \end{pmatrix}$$

and

$$\Sigma_k^* = \frac{\Delta t}{(M-k)} \begin{pmatrix} (M-k-1)\beta(X_k,\Theta) & (M-k-1)\beta(Y_k^{xz},\Theta) \\ (M-k-1)\beta(Y_k^{zx},\Theta) & C \end{pmatrix}$$

while $C = (M-k)\beta(Z_k,\Theta) - \beta(Y_k^{zx},\Theta)\beta(X_k,\Theta)^{-1}\beta(Y_k^{xz},\Theta)$. Moreover, we simplify the notations of $Y_k^{zx} = (Z_k, X_k)$ and $Y_k^{xz} = (X_k, Z_k)$ by Y_k^{zx} and Y_k^{xz}, respectively. By this approach, a diffusion bridge can be constructed starting from Y_j and finishing at x_M so that it can lie from Y_{j+1} to Y_{M-1} and Eq. 20 can be taken as the candidate generator of the states within the Metropolis-within-Gibbs algorithm. In these expressions, $\mu(.,\Theta)$ and $\beta(.,\Theta)$ describe the mean and the diffusion matrix of the associate components of Y given the reaction rates. Finally, the acceptance probability for the update of state is computed via the unnormalized form of the complete likelihood ratio as below.

$$\alpha(Y_{j+1},\ldots,Z_M,\ldots,Y_{M^+-1}|Y_{j+1}^*,\ldots,Z_M^*,\ldots,Y_{M^+-1}^*) = \min\left\{1, \frac{\prod_{k=j}^{M^+-1}\pi(Y_{k+1}^*|Y_k^*,\Theta)}{\prod_{k=j}^{M^+-1}\pi(Y_{k+1}|Y_k,\Theta)} \frac{q(Y_{j+1},\ldots,Z_M,\ldots,Y_{M^+-1}|Y_j,x_M,Y_{M^+},\Theta)}{q(Y_{j+1}^*,\ldots,Z_M^*,\ldots,Y_{M^+-1}^*|Y_j,x_M,Y_{M^+},\Theta)}\right\} \quad (21)$$

in which $q(.|.)$ displays the transition kernel density of the move. For the update of the proposal state, as presented previously, a standard uniform random variable u is generated and the proposal blocks of states is accepted if $u < \alpha(.|.)$ given in Eq. 21 is satisfied. Otherwise, the current state is accepted. Lastly, the underlying procedure is repeated until the convergence for all columns of Y is achieved.

On the other hand, for the estimation of the reaction rate constants Θ, we can still perform the same random sampling plan as described in Eqs. 17 and 18 under the Metropolis-within-Gibbs by columnwise updates.

If the performance of the diffusion bridge method is compared with the results of the columnwise update of states that are both based on the Metropolis-within-Gibbs sampling, it is seen that the former is more successful in decreasing the high correlation between latent or augmented states. Therefore it can achieve better mixing based on the same number of MCMC iterations. In Figs. 1 and 2, we present the plots of the MCMC runs based on the same number of iterations computed by the columnwise and the modified diffusion bridge updates, respectively, modelled by the diffusion approximation. The results indicate that the graphs of the modified diffusion bridge method reach its convergence faster than the columnise update, resulting in better mixing properties. The same conclusion is also found when the estimates of the conditional posteriors are compared with respect to their true values. The estimates of the modified diffusion bridge method have higher accuracy than the estimates of the columnwise updates.

However, as we keep the dependency between the model parameters Θ and the latent states, the convergence rate can alter with respect to m, the number of augmented states within every pair of observed time points, and the number of obser-

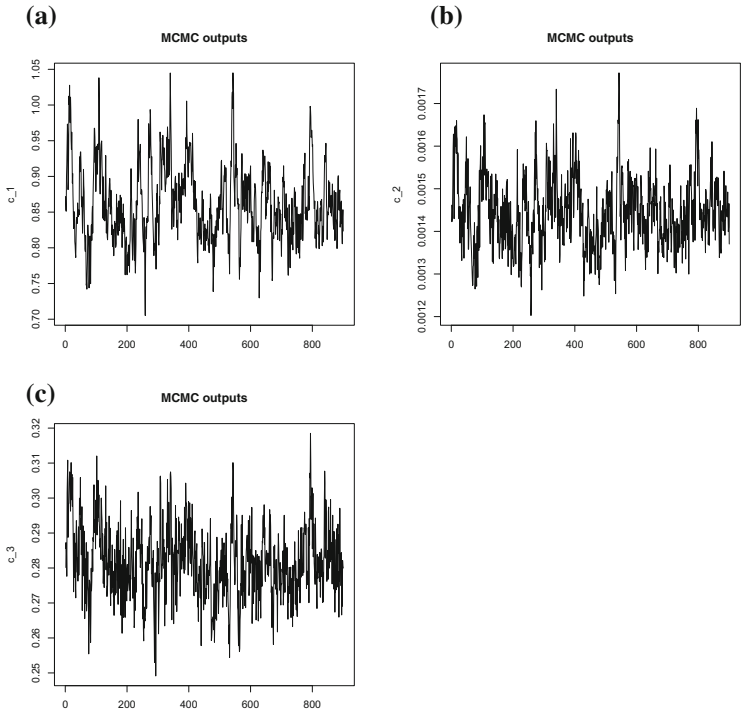


Fig. 1 The trace plots of the MCMC outputs based on 1,000,000 MCMC runs for a toy set modelled by the diffusion approximation and estimated by the Metropolis-within-Gibbs methods with columnwise updates

vations in Y. If even one of these values is high, the mixing becomes slower [14]. Hereby, in order to solve this source of dependency, there are two major alternative approaches. The first one is to use the reparametrization of the missing data whose algorithm is based on the irreducible MCMC (Markov Chain Monte Carlo) technique for all values of m [22]. Although this method gives promising results to overcome the dependency, its application is currently limited for univariate diffusion models [2, 16, 22]. Whereas, the second approach suggests to update both the states and the model parameters simultaneously and this approach is called as the particle filtering method whose mathematical details are presented in the following part.

3.3 Particle Filtering Method

The *particle filtering method* is an alternative approach to unravel the problem of dependency between the latent states Y and the model parameters Θ in the diffusion approximation. This method has been performed by [1, 4] for unobserved state

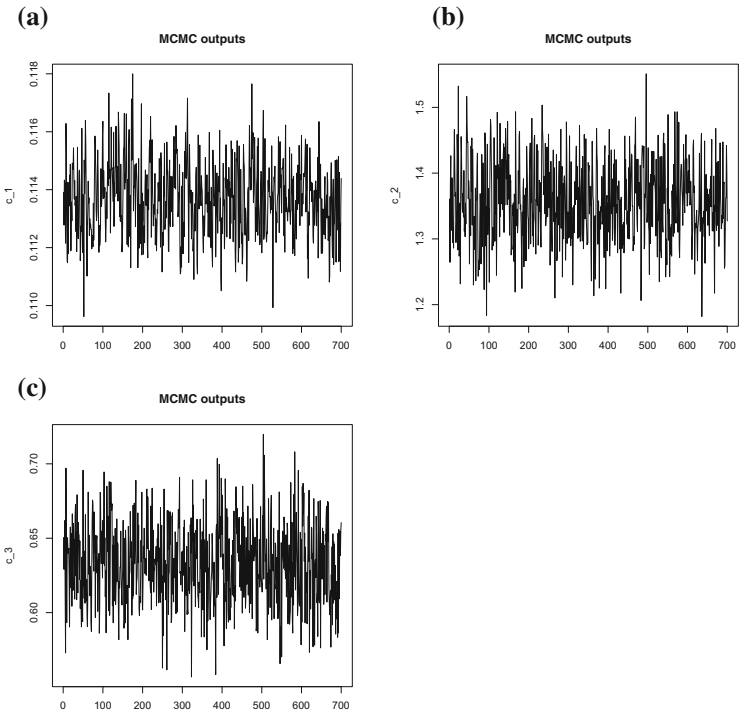


Fig. 2 The trace plots of the MCMC outputs based on 1,000,000 MCMC runs for a toy set modelled by the diffusion approximation and estimated by the Metropolis-within-Gibbs methods with the modified diffusion bridge method

variables in discrete time-course data. Then [14, 15] propose to implement it for the partially observed and augmented datasets. In this approach, in addition of the modified diffusion bridge method, the update of latent states is conducted with the update of the partially observed states $Z_{j+1}, \ldots, Z_{M-1}, Z_M$ and Θ simultaneously. Here, j denotes the integer multiple of m augmented states. Therefore each observed value, denoted by x_M $(M = j + m)$, is seen at every t_M after $(m - 1)$ latent states, i.e., $X_{j+1}, \ldots, X_{M-1}$. Accordingly, to generate a joint candidate of Θ and Z_j at time t_M for observed x_M, i.e., $\pi_M(\Theta, Z_j)$, we can apply their joint posterior as shown in Eq. 22 within the time interval $[t_j, t_M)$.

$$\pi_M(\Theta, Z_M) \propto \pi_j(\Theta, Z_j) \prod_{k=j}^{M-1} \pi(Y_{k+1}|Y_k, \Theta) \tag{22}$$

in which $Z_j, Y_{j+1}, \ldots, Y_{M-1}$ are integrated out from this target distribution. Whereas, as the first term on the right hand side in Eq. 22 can not be derived explicitly, it is

approximated by the set of points or *particles* $(\Theta_{(s)}, Z_{j(s)})$ where $s = 1, \ldots, r$ and r is the total number of reaction rates. Hence, $\Theta_{(s)}$ presents the s-dimensional block update of Θ when originally each $\Theta = (c_1, \ldots, c_r)$ has such a vectorial form. In this block update, each $(\Theta_{(s)}, Z_{j(s)})$ pair has a $1/r$ constant probability of selection and when $r \to \infty$, the particles can approximate the target density of $\pi_j(\Theta_j, Z_j)$ truly. Therefore, to produce this density, we can use different approaches. Here, we perform the *simulation filter algorithm* which is based on the MCMC sampling of $(Z_j, Y_{j+1}, \ldots, Y_{M-1}, Z_M, \Theta)$ and marginalizing only (Θ, Z_M) [14, 15]. In the simulation filter algorithm, the proposal value of (Θ^*, Z_j^*) is produced from the multivariate normal distribution as presented in Eq. 23.

$$(\Theta^*, Z_j^*) \sim N\left\{(\Theta_{(u)}, Z_{j(u)})', \gamma^2\psi\right\}. \tag{23}$$

In the above expression, u stands for an integer selected from 1 to r with equal probability $1/r$. Furthermore, γ^2 means the *smoothing parameter* that can be thought like the tuning parameter used in the update of the model parameters in all updating schemes. ψ represents the Monte Carlo posterior variance.

Accordingly, if we generate candidate Θ and Z_j, (Θ^*, Z_j^*) from Eq. 23, the candidate latent states, $(Y_{j+1}^*, \ldots, Y_{M-1}^*, Z_M^*)$ is produced by using the modified diffusion bridge method whose proposal is found by Eq. 20. Then for given Y_{M-1}^*, x_M and Θ^*, we can simulate Z_M^* from $Z_M^* \sim \pi(Z_M^*|Y_{M-1}^*, x_M, \Theta)$ from the following conditional proposal.

$$\pi(Y_{k+1}|Y_{M^+}, Y_k, x_M, \Theta) \sim N(Y_k + \mu_k^{**}, \Sigma_k^{**}) \tag{24}$$

for

$$\mu_k^{**} = \mu(Y_k, \Theta)\Delta t + \Delta t(\beta(Y_k, \Theta), C_k) \times \Lambda^{-1} \times \begin{pmatrix} Y_{M^+} - [Y_k + \mu(X_k, \Theta)\Delta^{\sim}] \\ x_M - [X_k + \mu(X_k, \Theta)\Delta^{\sim}] \end{pmatrix}$$

and

$$\Sigma_k^{**} = \beta(Y_k, \Theta)\Delta t - \Delta t(\beta(Y_k, \Theta), C_k) \times \Lambda^{-1} \times \begin{pmatrix} \beta(Y_k, \Theta) \\ C_k' \end{pmatrix} \times \Delta t$$

where $C_k = (\beta(X_k, \Theta), \beta(Y_k^{xz}, \Theta))'$. Here, $(.)'$ denotes the transpose of the given term and $\Delta^{\sim} = (M^+ - k)\Delta t$. Finally,

$$\Lambda = \begin{pmatrix} \beta(Y_k, \Theta)\Delta^{\sim} & C_k\Delta t \\ C_k'\Delta t & \beta(X_k, \Theta)\Delta^{\sim} \end{pmatrix}.$$

In the end, the algorithm updates the system by the acceptance probability α given below.

$$\alpha = \frac{\prod_{k=j}^{M-1} \pi(Y_{k+1}^*|Y_k^*, \Theta)}{\prod_{k=j}^{M-1} \pi(Y_{k+1}|Y_k, \Theta)} \times \frac{\pi(Z_M|Y_{m-1}, x_M, \Theta) \prod_{k=j}^{M-2} \widetilde{\pi}(Y_{k+1}|Y_k, x_M, \Theta)}{\pi(Z_M^*|Y_{M-1}^*, x^M, \Theta^*) \prod_{k=j}^{M-2} \widetilde{\pi}(Y_{k+1}^*|Y_k^*, x_M, \Theta^*)} \quad (25)$$

in which $\widetilde{\pi}(.|.)$ terms on the right hand side in Eq. 25 indicates the transition kernel density presented in Eq. 20. Here, the move is accepted by putting $\Theta_{(s)} = \Theta_{(s)}^*$ and $Y_{(s)} = Y_{(s)}^*$ $(s = 1, \ldots, r)$ with a probability $\min(1, \alpha)$. Lastly, similar to previous updating rules, the algorithm repeats this calculation for all $j := j + m$ iteratively until the convergence is obtained for both Θ and Y.

In general, from comparative analyses of this approach with the columnwise updates and the modified diffusion bridge methods, the particle filtering algorithm is more successful in decreasing the severity of the dependency between the model parameters and the latent states in inference. Therefore, it can propose a better mixing properties in the MCMC scheme [14, 15].

4 Conclusion

In this chapter, we have presented an overview of methods for the stochastic modelling and inference of the model parameters, i.e., stochastic reaction rate constants, for the realistically complex biological systems. Among alternatives, we have explained in details the diffusion approximation methods for modelling and the estimation of the model parameters via this model. In the description of the inference approaches, we have followed a plan from the less complex to more sophisticated approaches in order to overcome the problem of dependency in the updates of partially observed states and reaction rates. All the suggested methods are based on the fully Bayesian inference by using the time-course dataset and the given inference techniques are very general in the sense that they can be applicable for any other stochastic modelling approaches with time-course data. Whereas, they are typically very computationally demanding. But, the underlying computational cost can significantly decrease if we can work on fully observed, rather than partially observed, systems. In both cases, we consider that regarding the detailed information that they can provide about the system, they are still worthy to perform.

Acknowledgements The author would like to thank to the grant of TUBITAK (Project No: TBAG: 112T772) for their financial support.

References

1. Berzuini, C., Best, N.G., Gilks, W.R., Larizza, C.: Dynamic conditional independence models and Markov Chain Monte Carlo methods. J. Am. Stat. Assoc. **92**(440), 1403–1412 (1997)
2. Beskos, A., Papaspiliopoulos, O., Roberts, G., Fearnhead, P.: Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. J. Roy. Stat. Soc. B **68**, 1–29 (2006)
3. Bower, J.M., Bolouri, H.: Computational Modelling of Genetic and Biochemical Networks, 2nd edn. Massachusetts Institute of Technology, Cambridge (2001)
4. Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. **10**, 197–208 (2000)
5. Durham, G.B., Gallant, A.R.: Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. J. Bus. Econ. Stat. **20**(3), 297–338 (2002)
6. Eraker, B.: MCMC analysis of diffusion models with application to finance. J. Bus. Econ. Stat. **19**(2), 177–191 (2001)
7. Gamerman, D., Lopes, H.F.: Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall/CRC, Boca Raton (2006)
8. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton (2004)
9. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977)
10. Gillespie, D.T.: The multivariate langevin and fokker-planck equations. Am. J. Phys. **64**(10), 1246–1257 (1996)
11. Gillespie, D.T.: The chemical langevin equation. J. Chem. Phys. **113**, 97–306 (2000)
12. Gillespie, D.T.: Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. **115**, 1716–1733 (2001)
13. Golightly, A., Wilkinson, D.J.: Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics **61**(3), 781–788 (2005)
14. Golightly, A., Wilkinson, D.J.: Bayesian sequential inference for nonlinear multivariate diffusions. Stat. Comput. **16**, 323–338 (2006)
15. Golightly, A., Wilkinson, D.J.: Bayesian sequential inference for stochastic kinetic biochemical network models. J. Comput. Biol. **13**(3), 838–851 (2006)
16. Golightly, A., Wilkinson, D.J.: Bayesian inference for nonlinear multivariate diffusion models observed with error. Comput. Stat. Data Anal. **52**(3), 1674–1693 (2008)
17. Jong, H.D.: Modelling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. **9**(1), 67–103 (2002)
18. Pedersen, A.R.: A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand. J. Stat. **2**(1), 55–71 (1995)
19. Purutçuoğlu, V., Wit, E.: Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Anal. **3**(4), 851–886 (2008)
20. Purutçuoğlu, V.: Inference of the stochastic MAPK pathway by modified diffusion bridge method. Cent. Eur. J. Oper. Res., pp. 1–15 (2012). https://doi.org/10.1007/s10100-012-0237-8
21. Risken, H.: The Fokker-planck Equation: Methods of Solution and Applications. Springer, Berlin (1989)
22. Roberts, G.O., Stramer, O.: On inference for partially observed nonlinear diffusion models using the metropolis-hastings algorithm. Biometrika **88**(3), 603–621 (2001)
23. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk metropolis algorithms. Ann. Appl. Probab. **7**(1), 110–120 (1997)
24. Turner, T.E., Schnell, S., Burrage, K.: Stochastic approaches for modelling in vivo reactions. Comput. Biol. Chem. **28**, 165–178 (2004)
25. Van Kampen, N.G.: Stochastic Processes in Physics and Chemistry. North- Holland, Amsterdam (1981)
26. Wilkinson, D.J.: Stochastic Modelling for Systems Biology. Chapman and Hall/CRC, Boca Raton (2006)

Complete Nonholonomy of the Rolling Ellipsoid - A Constructive Proof

F. Rüppel, F. Silva Leite and R. C. Rodrigues

Abstract We present a constructive proof of the complete nonholonomy of the rolling ellipsoid. The rolling motions are assumed to be over the affine tangent space at a point of the n-ellipsoid and both manifolds are considered embedded in $\mathbb{R}^{n+1}$, equipped with a metric that results from a convenient deformation of the Euclidean metric. The deformation is defined through a positive definite matrix D whose eigenvalues are the semi-axis of the ellipsoid. The rolling motion has the usual constraints of non-slip and non-twist. Showing that the rolling ellipsoid is a complete nonholonomic system reduces to showing that one can move between two arbitrary admissible configurations by rolling without slipping and without twisting. We exhibit piecewise linear paths on the affine tangent space along which the ellipsoid rolls in order to perform the forbidden motions, twists and slips.

Keywords Ellipsoid · Group of isometries · Rolling maps · Twist · Slip
Nonholonomic constraints · Kinematic equations · Controllability

F. Rüppel
Institute of Mathematics, University of Würzburg,
Emil-Fischer-Straße 40, 97074 Würzburg, Germany
e-mail: frederike.rueppel@mathematik.uni-wuerzburg.de

F. Silva Leite (✉)
Institute of Systems and Robotics, University of Coimbra,
Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
e-mail: fleite@mat.uc.pt

F. Silva Leite
Department of Mathematics, University of Coimbra,
Largo D. Dinis, 3001-454 Coimbra, Portugal

R. C. Rodrigues
Department of Physics and Mathematics, ISEC - Polytechnic Institute of Coimbra,
Rua Pedro Nunes, 3030-199 Coimbra, Portugal
e-mail: ruicr@isec.pt

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_19

1 Introduction

This paper deals with controllability of a rolling system consisting of an ellipsoid rolling on the affine tangent space at a point. This is a particular situation of a manifold rolling, without slipping and twisting, on another manifold of equal dimension, so that both manifolds are tangent at every point of contact. The first manifold is moving (rolling manifold) while the second is static. By an admissible configuration of such rolling system we mean any position of the rolling manifold in which it is tangent to the static manifold. It is well known that the most classical of all rolling systems, consisting of a sphere rolling over its tangent plane at a point, always with the constraints of no-twist and no-slip, is a complete nonholonomic system, which is the same as saying that the system is controllable. In other words, given any two admissible configurations of the sphere, it is always possible to make it roll over the tangent plane from one configuration to the other, without violating the constraints. This controllability issue has been studied for other particular rolling motions and even for general rolling manifolds, such as in [2]. These results, however, do not tell us how to do it. For the Euclidean 2−sphere this problem has been studied before in [1, 5], and the results have been extended to the Euclidean n-sphere in [6]. The objective of this article is to present a constructive proof for the controllability of the rolling motion of the n-dimensional ellipsoid over the affine tangent space at a point, assuming that both manifolds are embedded in $\mathbb{R}^{n+1}$ equipped with a left invariant metric that results from a deformation of the Euclidean metric. This will be achieved by showing that the forbidden motions of twist and slip can be alternatively performed by rolling without twist and without slip. The 2-dimensional case serves as an inspiration and illustration. Before we reach the main results in Sect. 6, we introduce the right geometry of the ellipsoid, the group of isometries of the embedding space, and the kinematic equations for rolling the ellipsoid over the affine tangent space at a point. These kinematic equations can be solved explicitly when rolling along geodesics. This fact is particularly important and will be used later in order to define the rolling path along which the ellipsoid has to roll in order to perform the forbidden motions of twist and slip. Throughout the paper we deal with the n-dimensional case, but the details and illustrations are presented for the case $n = 2$.

2 Geometry of the Ellipsoid

Let $d_1, d_2, \ldots, d_{n+1}$ be positive real numbers. The n-dimensional ellipsoid associated to these numbers is defined as

$$\mathcal{E}^n := \left\{ (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : \frac{x_1^2}{d_1^2} + \cdots + \frac{x_{n+1}^2}{d_{n+1}^2} = 1 \right\}. \tag{1}$$

The positive definite matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_{n+1}) \succ 0$ induces a metric on $\mathbb{R}^{n+1}$ defined by

$$(U, V) \mapsto \langle U, V \rangle_{D^{-2}} := \langle U, D^{-2}V \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean metric in $\mathbb{R}^{n+1}$. This metric space will be denoted by $M = (\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle_{D^{-2}})$. Since

$$\langle DU, DV \rangle_{D^{-2}} = \langle DU, D^{-2}DV \rangle = \langle U, V \rangle,$$

the mapping

$$\begin{array}{rccc} \varphi\colon & (\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle) & \to & M \\ & x & \mapsto & Dx \end{array}, \tag{3}$$

is an isometry. M is an example of a space equipped with a left-invariant metric. Unlike the Euclidean space, groups of isometries acting on objects from the left hand side is different than from the right. The main reason for introducing M is that when the ellipsoid is embedded in Euclidean space even curves with a simple geometry, such as geodesics, are very hard to compute. To avoid this problem, we follow what has been done in [7, 8] and embed the ellipsoid $\mathcal{E}^n$ in the Riemannian manifold $M = \left(\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle_{D^{-2}}\right)$ with $\langle U, V \rangle_{D^{-2}} := \langle U, D^{-2}V \rangle$, so that the ellipsoid $\mathcal{E}^n$ behaves like the sphere in Euclidean space with its standard metric. With respect to this Riemannian metric, the corresponding geodesics can be expressed in closed form.

From now on, we consider the n-dimensional ellipsoid $\mathcal{E}^n$ and its affine tangent space $V \cong T^{\mathrm{aff}}_{p_0}\mathcal{E}^n$ at a particular point $p_0 \in \mathcal{E}^n$, both embedded in $M = \left(\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle_{D^{-2}}\right)$. The ellipsoid $\mathcal{E}^n$ is allowed to roll along any path in V without twisting and without slipping. To describe the rolling motions we need the action on M of its group of isometries.

3 Group of Isometries of $M = \left(\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle_{D^{-2}}\right)$

Let Isom (M) denote the (Lie) group of isometries of M. If $\varphi\colon M \to M$ is an isometry, then, for any $p \in M$ and $U, V \in T_pM$, the following equality holds, where φ_* is the push-forward of φ.

$$\langle U, V \rangle_{D^{-2}} = \langle \varphi_* U, \varphi_* V \rangle_{D^{-2}}$$

or, equivalently,

$$\langle U, D^{-2}V \rangle = \langle \varphi_* U, D^{-2}\varphi_* V \rangle = \langle U, \varphi_*^T D^{-2} \varphi_* V \rangle.$$

So, it follows that $D^{-2} = \varphi_*^T D^{-2} \varphi_*$, that is, $\varphi_* \in \mathcal{G}_{D^{-2}}$, where $\mathcal{G}_{D^{-2}}$ is the matrix quadratic Lie group defined as

$$\mathcal{G}_{D^{-2}} := \left\{ X \in GL(n+1) : \; X^T D^{-2} X = D^{-2} \right\}. \tag{4}$$

The Lie algebra of $\mathcal{G}_{D^{-2}}$ is defined as:

$$\mathcal{L}_{D^{-2}} := \{ A \in \mathfrak{gl}(n+1) : \; A^T D^{-2} = -D^{-2} A \}. \tag{5}$$

It can be easily seen that, for any $g \in \mathcal{G}_{D^{-2}}$, there exists exactly one $R \in \mathbb{SO}(n+1)$ such that $g = DRD^{-1}$. Therefore $\mathcal{G}_{D^{-2}} = D\,\mathbb{SO}(n+1)\,D^{-1}$ and so the two groups are isomorphic ($\mathcal{G}_{D^{-2}} \cong \mathbb{SO}(n+1)$). Also, for any $\Omega \in \mathcal{L}_{D^{-2}}$, there exists exactly one $A \in \mathfrak{so}(n+1)$ such that $\Omega = DAD^{-1}$.

The Lie group of isometries of M is then

$$\mathrm{Isom}(M) = \mathcal{G}_{D^{-2}} \ltimes \mathbb{R}^{n+1} \cong \mathbb{SE}(n+1). \tag{6}$$

In the reminder of this paper the elements of $\mathrm{Isom}(M)$ will be denoted, as usual, by pairs (X, s), with $X \in \mathcal{G}_{D^{-2}}$ and $s \in \mathbb{R}^{n+1}$. The group operations in $\mathcal{G}_{D^{-2}} \ltimes \mathbb{R}^{n+1}$ are defined as: $(X, s)^{-1} = \left(X^{-1}, -X^{-1}s\right)$ and $(X_1, s_1) \cdot (X_2, s_2) = (X_1 X_2, X_1 s_2 + s_1)$.

4 Rolling Maps

Rolling maps describe how two manifolds of the same dimension, both embedded in a Riemannian space, roll on each other without twisting and without slipping. Such a motion will be sometimes referred as "pure rolling". A rolling map is a curve in the group of isometries of the embedded manifold that satisfies certain constraints. The first formal definition of a rolling map for Euclidean submanifolds appeared in Sharpe [10], but in the meanwhile it has been extended to more general Riemannian manifolds in [4]. In a general situation, there are two no-twist conditions (tangential and normal conditions). But when the rolling manifold has co-dimension one, the normal condition is always satisfied. We recall below the definition of rolling map in the co-dimension one situation, since this is the particular case we are about to study. In what follows $I \subset \mathbb{R}$ denotes a closed interval and $N^\perp$ denotes the orthogonal space of an embedded submanifold $N \subset M$, with respect to the Riemannian metric on M.

Definition 1 Let M_0 and M_1 be two n-manifolds isometrically embedded in an $n+1$-dimensional Riemannian manifold M and $\sigma_1 \colon I \to M_1$ a piecewise smooth curve in M_1. A *rolling map* of M_1 on M_0, without slipping or twisting, is a (piecewise smooth) map $\chi \colon I \to \mathrm{Isom}(M)$ satisfying the following conditions:

Rolling Conditions:

There exists a piecewise smooth curve $\sigma_1 : I : M_1$ (called the rolling curve) satisfying:

1. $\chi(t)(\sigma_1(t)) \in M_0$, for all $t \in I$;
2. $T_{\chi(t)(\sigma_1(t))}(\chi(t)(M_1)) = T_{\chi(t)(\sigma_1(t))}M_0$, for all $t \in I$.

The curve $\sigma_0 \colon I \to M_0$ defined by $\sigma_0(t) := \chi(t)(\sigma_1(t))$ is called the *development curve* of σ_1.

No-slip Condition:

$\dot{\sigma}_0(t) = \chi(t)_*(\dot{\sigma}_1(t))$, for almost all $t \in I$;

No-twist Condition:

$\left(\dot{\chi}(t) \circ \chi(t)^{-1}\right)_* (T_{\sigma_0(t)}M_0) \subset \left(T_{\sigma_0(t)}M_0\right)^{\perp}$, for almost all $t \in I$.

■

The no-slip and no-twist conditions are nonholonomic conditions, i.e. nonintegrable constraints on velocities that must hold for any $t \in I$ for which the derivatives are defined. For rolling maps of general manifolds there is a second no twist condition where the roles of $T_{\sigma_0(t)}M_0$ and $\left(T_{\sigma_0(t)}M_0\right)^{\perp}$ appear interchanged.

Remark 1 Using this definition and properties of isometries, one can easily prove that rolling is invariant under the action of any isometry of the embedding space. More precisely, if M_1 rolls on M_0 with rolling map χ, rolling curve σ_1, and development curve σ_0, and if $\chi_f \in \mathrm{Isom}(M)$ is a fixed isometry, then $\chi_f(M_1)$ rolls on $\chi_f(M_0)$ with rolling map $\chi_f \cdot \chi \cdot {\chi_f}^{-1}$, rolling curve $\chi_f(\sigma_1)$ and development curve $\chi_f(\sigma_0)$.

5 Kinematic Equations for the Rolling Ellipsoid

The kinematic equations for the rolling motion of the ellipsoid over the affine tangent space at a point have been derived in [7], precisely for the situation when both manifolds have the metric induced by $< ., . >_{D^{-2}}$. When the rolling motion is performed along geodesics or broken geodesics and, consequently, the development curves are piecewise linear, the kinematic equations are easy to solve. This is the situation that we will explore later, and for that reason we include next the essentials about kinematic equations taken from [7], when the point of contact is $p_0 = (0, \cdots, 0, -d_3)^{\top}$. In this case, the skew-symmetric matrix $A(t) = u(t){p_0}^{\top}D^{-1} - D^{-1}p_0u(t)^{\top}$, with $u = (u_1, \cdots, u_n, 0)^{\top}$, which appear in the next proposition, has the following structure:

$$A(t) = \begin{pmatrix} 0 & \dots & 0 & -u_1(t) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & -u_n(t) \\ u_1(t) & \dots & u_n(t) & 0 \end{pmatrix}. \tag{7}$$

Proposition 1 ([7], Proposition 5.1) *Let* $X\colon I \to \mathcal{G}_{D^{-2}}$ *and* $s\colon I \to \mathbb{R}^{n+1}$ *be the solution of the following set of equations*

$$\begin{cases} \dot{s}(t) = -DA(t)D^{-1}p_0 \\ \\ \dot{X}(t) = DA(t)D^{-1}X(t) \end{cases} \tag{8}$$

satisfying $s(0) = 0$ *and* $X(0) = \mathrm{id}$, *where* $A(t)$ *is given by (7).*

Then, $\chi = (X, s)\colon I \to G_{D^{-2}} \ltimes \mathbb{R}^{n+1}$ *is the rolling map of the rolling motion of the ellipsoid* $\mathcal{E}^n$ *over its affine tangent space at* p_0, *with rolling curve* $\sigma_1(t) = (X(t))^{-1}p_0$ *and development curve* $\sigma_0(t) = s(t) + p_0$.

Corollary 1 ([7], Corollary 5.2) *If* $A(t) = A$ *is a constant matrix, the solution of the kinematic equations (8) is given by*

$$\begin{cases} s(t) = -tDAD^{-1}p_0 \\ \\ X(t) = D\exp(tA)D^{-1} \end{cases}, \tag{9}$$

and the rolling curve and its development, given respectively by

$$\begin{aligned} \sigma_1(t) &= X_0^{-1}D\exp(-tA)D^{-1}p_0, \\ \sigma_0(t) &= -tDAD^{-1}p_0 + p_0, \end{aligned} \tag{10}$$

are geodesics on the ellipsoid $\mathcal{E}^n$ *and on its affine tangent space* $T_{p_0}^{aff}\mathcal{E}^n$, *with respect to the metric* $\langle\cdot,\cdot\rangle_{D^{-2}}$.

Remark 2 It is clear from here that the rolling map $t \mapsto (X(t), s(t))$ is uniquely determined if one chooses the development curve $t \mapsto \sigma_0(t)$.

It has been proven in [9] that, as a consequence of the positivity of the Gaussian curvature of the ellipsoid, the rolling system consisting of the ellipsoid rolling on its affine tangent space at a point is completely nonholonomic (or controllable). That is, one can steer the ellipsoid from an admissible configuration (any configuration in which the ellipsoid is tangent to the affine tangent space) to any other admissible configuration, only by rolling without twist and without slip, that is by pure rolling only. Interested reader is referred to [9] for more details. The interesting issue now is to know how to do it. To answer this question, is it enough to show how to generate the forbidden motions (twists and slips) using pure rolling only. The main objective of this article is to address this problem. This will be done in the next section, starting with the 2-dimensional case. There is a strong analogy between what we do here and the ideas contained in [6], which in turn were inspired by [5].

6 Realizing Twists and Slips by Pure Rolling

Let e_i, $i = 1, 2, 3$ denote the standard basis vectors of $\mathbb{R}^3$, here represented as row matrices. Without loss of generality we might assume that p_0 is the "south pole" of $\mathcal{E}^2$, that is, $p_0 = -d_3 e_3^\top$. Otherwise, we apply a $\mathcal{G}_{D^{-2}}$ transformation to the ellipsoid to bring the point p_0 to that position.

When $p_0 = -d_3 e_3^\top$, a twist is an element of $\mathcal{G}_{D^{-2}}$ that fixes $e_3^\top$, while a slip is a translation. The figures below illustrate the forbidden motions (twist and slip) (Figs. 1 and 2).

We will exhibit a piecewise linear closed path that the ellipsoid has to trace on $V \cong T^{\text{aff}}_{p_0}\mathcal{E}^2$ in order to realize a twist, and a piecewise linear path that the ellipsoid has to trace on $V \cong T^{\text{aff}}_{p_0}\mathcal{E}^2$ in order to realize a slip.

First, we introduce the following basis of the Lie algebra of $\mathcal{G}_{D^{-2}}$, where $A_{i,j} := e_i e_j^\top - e_j e_i^\top$.

$$\left\{A^D_{ij} := De_i e_j^\top D^{-1} - De_j e_i^\top D^{-1} = DA_{i,j}D^{-1}, \quad 1 \le i < j \le 3\right\}. \tag{11}$$

Since $\mathcal{G}_{D^{-2}} = D\mathbb{SO}(3)D^{-1}$, it is immediate to conclude that a clockwise twist by an angle φ is given by

$$z_D(\varphi) := \mathrm{e}^{-\varphi A^D_{1,2}} = D \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} D^{-1}. \tag{12}$$

Analogously, we can define the following elements in $\mathcal{G}_{D^{-2}}$ that leave invariant either e_2 or e_1:

Fig. 1 A forbidden motion: a twist around the vertical axis

Fig. 2 A forbidden motion: a slip

$$
\begin{aligned}
y_D(\varphi) \quad &:= \mathrm{e}^{\varphi A^D_{1,3}} = D \begin{bmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{bmatrix} D^{-1}, \\
x_D(\varphi) &:= \mathrm{e}^{-\varphi A^D_{2,3}} = D \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{bmatrix} D^{-1}.
\end{aligned} \tag{13}
$$

Defining $\mathrm{ad}_A\, B := [A, B] = AB - BA$ and, recursively, $\mathrm{ad}^j_A\, B := \mathrm{ad}^{j-1}_A [A, B]$, for $j = 2, 3, \cdots$, one has

$$
\mathrm{e}^{tA}\, B\, \mathrm{e}^{-tA} = \mathrm{e}^{t\mathrm{ad}_A}\, B = B + t[A, B] + \cdots + \frac{t^k}{k!}\, \mathrm{ad}^k_A\, B + \cdots,
$$

Also, taking into consideration the following identities

$$
\left[A^D_{1,2}, A^D_{2,3}\right] = A^D_{1,3}, \quad \left[A^D_{1,2}, A^D_{1,3}\right] = -A^D_{2,3}, \quad \left[A^D_{1,3}, A^D_{2,3}\right] = -A^D_{1,2},
$$

it is straightforward to conclude that

$$
z_D(\varphi) = x_D(\pm\tfrac{\pi}{2})\, y_D(\pm\varphi)\, x_D(\mp\tfrac{\pi}{2}).
$$

Therefore, decomposing $z_D(\varphi)$ as $z_D(\varphi) = z_D(\frac{\varphi}{2})\, z_D(\frac{\varphi}{2})$, it follows that any twist can be written as:

$$
z_D(\varphi) = x_D(-\tfrac{\pi}{2})\, y_D(-\tfrac{\varphi}{2})\, x_D(\pi)\, y_D(\tfrac{\varphi}{2})\, x_D(-\tfrac{\pi}{2}). \tag{14}
$$

Notice that the sum of the arguments for x_D, and also for y_D, in the last expression add up to zero. We use this decomposition of $z_D(\varphi)$ to guarantee that the ellipsoid will return to the initial position (with p_0 being the point of contact with V).

6.1 Realizing a Twist

Next, we exhibit a closed piecewise linear path $\alpha_0(t)$, which is the development curve on V for the rolling motion (without twisting and without slipping) of the ellipsoid which realizes a clockwise twist of angle φ. This path is composed of five line segments which form a rectangle in the tangent plane. As mentioned in Remark 2, the whole rolling motion is determined by this choice of the development curve. Figure 3 illustrates the realization of such twist, showing the rolling ellipsoid at six different times, in particular, when it is touching the tangent plane V at the vertices of the rectangle. For the sake of clarity, the picture shows two rectangles, but in reality they overlap.

This rolling motion occurs in the time interval $[0, T]$, with $T = \pi d_2 + \varphi d_1$, partitioned as $0 = t_0 < t_1 < t_2 < t_3 < t_4 < t_5 = T$, where

$$\begin{aligned} t_1 &= \frac{\pi}{2} d_2 \\ t_2 &= t_1 + \frac{\varphi}{2} d_1 = \frac{\pi}{2} d_2 + \frac{\varphi}{2} d_1 \\ t_3 &= t_2 + \pi d_2 = \frac{3\pi}{2} d_2 + \frac{\varphi}{2} d_1 \\ t_4 &= t_3 + \frac{\varphi}{2} d_1 = \frac{3\pi}{2} d_2 + \varphi d_1 \\ t_5 &= t_4 + \frac{\pi}{2} d_2 = \pi d_2 + \varphi d_1 \quad , \end{aligned} \tag{15}$$

and the development curve is defined by:

$$\sigma_0{}^{\top}(t) = \begin{cases} (0, t, -d_3)^{\top}, & 0 \le t \le t_1 \\ (t - \frac{\pi}{2} d_2, \frac{\pi}{2} d_2, -d_3)^{\top}, & t_1 < t \le t_2 \\ (\frac{\varphi}{2} d_1, -t + \frac{\varphi}{2} d_1 + \pi d_2, -d_3)^{\top}, & t_2 < t \le t_3 \\ (-t + \varphi d_1 + \frac{3\pi}{2} d_2, -d_3)^{\top}, & t_3 < t \le t_4 \\ (0, t - \varphi d_1 - 2\pi d_2, -d_3)^{\top}, & t_4 < t \le t_5 \end{cases} . \tag{16}$$

This is all we need to define the sequence of rolling motions along the broken geodesic that performs the twist. These rolling motions can be written in terms of x_D and y_D only. Next, we exhibit the equations of five manifolds, $N_1, \cdots, N_5$, the first four for the rolled ellipsoid when it is tangent to the plane at every corner of the rectangle, and N_5 which only differs from $N_0 =: \mathcal{E}^2$ by a twist of angle φ. We omit straightforward computations that enable the simplifications below. The six manifolds, N_i, $i = 0, \cdots, 5$, can be observed in Fig. 3, where the matrix that defines the ellipsoid was chosen to be $D = \mathrm{diag}\{4, 2, 1\}$. The sizes of all the surfaces N_i is the same, although due to perspective they appear to be different.

$$N_1 = x_D(-\pi/2)\mathcal{E}^2 + \begin{bmatrix} 0 \\ \frac{\pi}{2} d_2 \\ 0 \end{bmatrix};$$

$$N_2 = y_D(\varphi/2) \left(N_1 - \begin{bmatrix} 0 \\ \frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \right) + \begin{bmatrix} \frac{\varphi}{2} d_1 \\ \frac{\pi}{2} d_2 \\ 0 \end{bmatrix}$$

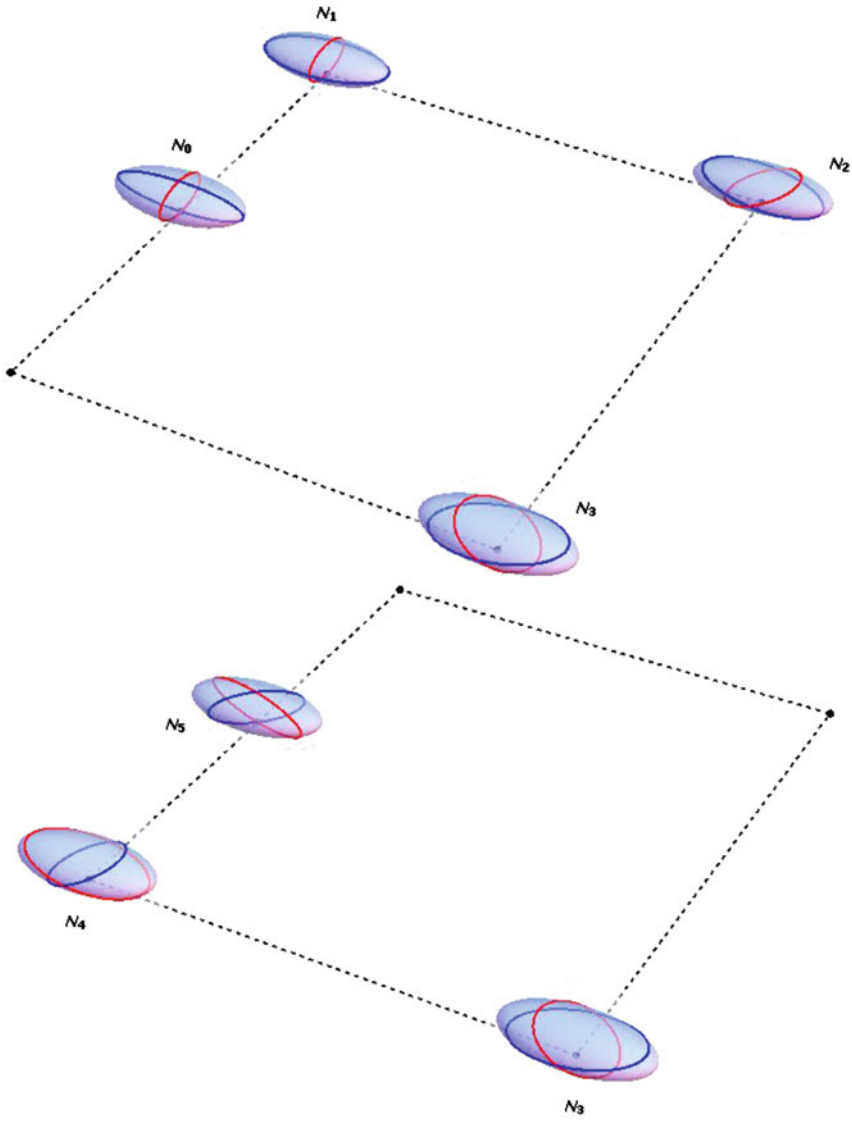


Fig. 3 Performing a twist, from N_0 to N_5

$$
\begin{aligned}
&= y_D(\varphi/2)\, x_D(-\pi/2)\, \mathcal{E}^2 + \begin{bmatrix} \frac{\varphi}{2} d_1 \\ \frac{\pi}{2} d_2 \\ 0 \end{bmatrix}; \\
N_3 &= x_D(\pi) \left(N_2 - \begin{bmatrix} \frac{\varphi}{2} d_1 \\ \frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \right) + \begin{bmatrix} \frac{\varphi}{2} d_1 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \\
&= x_D(\pi)\, y_D(\varphi/2)\, x_D(-\pi/2)\, \mathcal{E}^2 + \begin{bmatrix} \frac{\varphi}{2} d_1 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix}; \qquad (17) \\
N_4 &= y_D(-\frac{\varphi}{2}) \left(N_3 - \begin{bmatrix} \frac{\varphi}{2} d_1 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \right) + \begin{bmatrix} 0 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \\
&= y_D(-\frac{\varphi}{2})\, x_D(\pi)\, y_D(\varphi/2)\, x_D(-\pi/2)\, \mathcal{E}^2 + \begin{bmatrix} 0 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix}; \\
N_5 &= x_D(-\frac{\pi}{2}) \left(N_4 - \begin{bmatrix} 0 \\ -\frac{\pi}{2} d_2 \\ 0 \end{bmatrix} \right) + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \\
&= x_D(-\frac{\pi}{2})\, y_D(-\frac{\varphi}{2})\, x_D(\pi)\, y_D(\varphi/2)\, x_D(-\pi/2)\, \mathcal{E}^2.
\end{aligned}
$$

So, using (14), we conclude that

$$N_5 = z_D(\varphi)\, \mathcal{E}^2, \qquad (18)$$

that is, the rolling motion (without twist and without slip) of the ellipsoid along the piecewise path given by (16) performs the required twist.

6.2 *Realizing a Slip*

A slip is a pure translation $s(\tau)$ that takes a submanifold N to $N + s(\tau)$, so without changing the orientation of N. The objective here is to show how to realize a slip from a point p_0 to another point p_2 in the tangent plane. Similarly to what has been done in [1] for the Euclidean sphere, we will show that a slip can be realized with pure rolling along at most two geodesic arcs.

Without loss of generality we assume that $N_0 := \mathcal{E}^2$, $p_0 = (0, 0, -d_3)^\top$, $p_2 = (x_2, 0, -d_3)^\top$, for $x_2 > 0$. In this case, $\tau = x_2$. While these assumptions simplify

calculations, they do not destroy the essence of the problem. This is due to the invariance of the rolling under the action of isometries of the embedding space (Remark 1).

First Case: If $\mathrm{dist}_{D^{-2}}(p_0, p_2)$ is a multiple of 2π, that is, if there exists a positive integer k such that

$$\mathrm{dist}_{D^{-2}}(p_0, p_2) = \sqrt{\langle p_2 - p_0, p_2 - p_0\rangle_{D^{-2}}} = x_2 {d_1}^{-1} = 2\pi k, \tag{19}$$

then a simple calculation shows that the pure rolling motion from $t = 0$ to $t = \tau$, resulting from the action of the rolling map $(X(t), s(t)) = \left(y_D(t{d_1}^{-1}), (t, 0, 0)^\top\right)$ performs the slip. To see this, notice that knowing $s(t)$ we can use (9) to obtain the matrix A and, consequently, $X(t)$. In this case, $A = {d_1}^{-1} A_{1,3}$ and $X(t) = D\exp(tA)D^{-1} = y_D(t{d_1}^{-1})$. But using the assumption (19), we may write $X(\tau) = y_D(\tau {d_1}^{-1}) = y_D(x_2 {d_1}^{-1}) = y_D(2\pi k) = I_3$. So, $X(\tau)N_0 + s(\tau) = N_0 + s(\tau)$ as required.

Second Case: If $\mathrm{dist}_{D^{-2}}(p_0, p_2)$ is not a multiple of 2π, let l be any integer satisfying $\mathrm{dist}_{D^{-2}}(p_0, p_2) < 2\pi l$. Then, there exist points $p_1 \in V$ such that $\mathrm{dist}_{D^{-2}}(p_0, p_1) = \mathrm{dist}_{D^{-2}}(p_1, p_2) = 2\pi l$. These points are of the form $p_1 = (\frac{x_2}{2}, y_1, -d_3)^\top$, with $y_1 = \pm\sqrt{d_2(4\pi l^2 - {d_1}^{-2}x_2^2/4)}$. The points p_0, p_1 and p_2 form the vertices of an isosceles triangle. In this case, the slip from p_0 to p_2 is realized by pure rolling along a broken geodesic that develops as two sides of this triangle, the ones joining p_0 to p_1 and p_1 to p_2. This clearly solves the problem since the distance between each pair of points is a multiple of 2π. For the sake of completeness, we include here the two rolling maps involved, (s_1, X_1) and (s_2, X_2). Since

$$\sigma_0(t) = \begin{cases} (t, t\frac{2y_1}{x_2}, -d_3)^\top, & 0 \le t \le \frac{x_2}{2} \\ (t, -t\frac{2y_1}{x_2} + 2y_1, -d_3)^\top, & \frac{x_2}{2} < t \le x_2 \end{cases},$$

it then follows from (9) that

$$A_1 = \begin{bmatrix} 0 & 0 & {d_1}^{-1} \\ 0 & 0 & \frac{2y_1}{x_2}{d_2}^{-1} \\ -{d_1}^{-1} & -\frac{2y_1}{x_2}{d_2}^{-1} & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & {d_1}^{-1} \\ 0 & 0 & -\frac{2y_1}{x_2}{d_2}^{-1} \\ -{d_1}^{-1} & \frac{2y_1}{x_2}{d_2}^{-1} & 0 \end{bmatrix},$$

and, consequently,

$$s_1(t) = \begin{cases} (t, t\frac{2y_1}{x_2}, 0)^\top, & 0 \le t \le \frac{x_2}{2} \\ (t, -t\frac{2y_1}{x_2} + 2y_1, 0)^\top, & \frac{x_2}{2} < t \le x_2 \end{cases}.$$

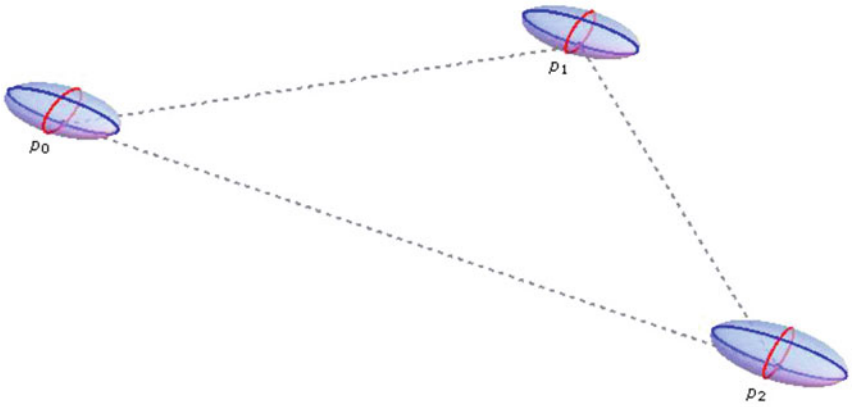


Fig. 4 Performing a slip, from p_0 to p_2

Rodrigues' formula for rotations in $\mathbb{R}^3$ tells us that if $A = a_1 A_{1,2} + a_2 A_{1,3} + a_3 A_{1,3}$, then

$$\exp(tA) = I_3 + \frac{A}{k}\sin(tk) + \frac{A^2}{k^2}(1 - \cos(tk)),$$

where $k = \sqrt{a_1^2 + a_2^2 + a_3^2}$. So, $X_1(t)$ and $X_2(t)$ can be easily obtained from A_1 and A_2 above.

Figure. 4 illustrates how to perform a slip. The sizes of the three surfaces are the same, although due to perspective they appear to be different.

6.3 A Glimpse at the General Case

For the n-dimensional case, the computations are more elaborate, but we can give a glimpse of the general ideas that are behind the realization of twists and slips. This is similar to what has been done in [6] for the Euclidean case and follows from the fact that the isometry group $\mathcal{G}_{D^{-2}}$ is related to the rotation group as $\mathcal{G}_{D^{-2}} = D\,\mathbb{SO}(n+1)\,D^{-1}$.

For the ellipsoid $\mathcal{E}^n$, let $p_0 = (0, \ldots, 0, -d_{n+1})^\top$. Consider the following basis for the Lie algebra of $\mathcal{G}_{D^{-2}}$:

$$\left\{A^D_{i,j} = D(e_i e_j^T - e_j e_i^T)D^{-1} = DA_{i,j}D^{-1},\ 1 \le i < j \le n+1\right\}.$$

A twist at p_0 is an element in $\mathcal{G}_{D^{-2}}$ that keeps p_0 invariant. It is easy to conclude that twists are the elements in the Lie subgroup of $\mathcal{G}_{D^{-2}}$ defined by $K := \exp \mathfrak{k}$, where

$$\mathfrak{k} := \operatorname{span}\{\mathrm{A}^{\mathrm{D}}_{\mathrm{i,j}},\ 1 \le \mathrm{i} < \mathrm{j} \le \mathrm{n}\} \cong \mathfrak{so}_{\mathrm{n}}.$$

Hence, the realization of a twist by pure rolling is equivalent to the possibility of expressing elements of $\exp \mathfrak{k}$ in terms of elements of $\exp \mathfrak{p}$, where $\mathfrak{p}$ is the vector space

$$\mathfrak{p} := \{\mathrm{A}^{\mathrm{D}}_{\mathrm{i,n+1}},\, 1 \leq \mathrm{i} \leq \mathrm{n}\}.$$

Note that, the Lie algebra $\mathcal{L}_{D^{-2}}$ of $\mathcal{G}_{D^{-2}}$ allows the direct sum decomposition

$$\mathcal{L}_{D^{-2}} = \mathfrak{k} \bigoplus \mathfrak{p},$$

which is a Cartan-like decomposition and induces a decomposition of $\mathcal{G}_{D^{-2}}$ as

$$\mathcal{G}_{D^{-2}} = KAK,$$

where $K = \exp \mathfrak{k}$ and $A = \exp \mathfrak{a}$, with $\mathfrak{a}$ a $1-$dimensional Abelian subalgebra contained in $\mathfrak{p}$. Details about these decompositions may be found in [3].

In order to be able to show how to realize a twist in higher dimensions, one needs to keep decomposing $\mathfrak{k}$ and K in a similar way until every element in K is written as a product of a finite number of elements from the one-parameter subgroups $\exp(\varphi A^D_{i,j})$, for $1 \leq i < j \leq n$. This is the counterpart to the decomposition of a rotation as a composition of Givens rotations, used in [6] for the case $D = I$.

Now, for each $A^D_{i,j} \in \mathfrak{k}$, there exist elements $A^D_{i,n+1}$ and $A^D_{j,n+1}$ in $\mathfrak{p}$ such that

$$\begin{aligned} &[A^D_{i,n+1}, A^D_{i,j}] = A^D_{j,n+1}, \quad [A^D_{i,j}, A^D_{j,n+1}] = A^D_{i,n+1}, \\ &[A^D_{j,n+1}, A^D_{i,n+1}] = A^D_{i,j}. \end{aligned} \tag{20}$$

And analogously to the $\mathcal{E}^2$ case, we may write

$$\mathrm{e}^{(\varphi A^D_{i,j})} = \mathrm{e}^{(\pm \frac{\pi}{2} A^D_{i,n+1})}\ \mathrm{e}^{(\pm \varphi A^D_{j,n+1})}\ \mathrm{e}^{(\mp \frac{\pi}{2} A^D_{i,n+1})}. \tag{21}$$

Proceeding as for $n = 2$, the result is the decomposition of every twist into a product of elements in $\exp \mathfrak{p}$, so that all rotation angles corresponding to a fixed axis $A^D_{i,n+1}$ add up to zero. So, any twist is realized by pure rolling along a closed broken geodesic.

Finally, due to the invariance of rolling with respect to the action of the isometry group of the embedding space (Remark 1), the realization of a slip from a point p_0 to a point p_2 in the affine tangent space at p_0 is similar to the case $n = 2$. So, we can assume, without loss of generality, that $p_0 = (0, \cdots, 0, -d_{n+1)})^\top$ and $p_2 = (x_2, \cdots, 0, -d_{n+1)})^\top$, for some $x_2 > 0$, and repeat the procedure in Sect. 6.2 just with the obvious adaptations. More precisely, if $\mathrm{dist}_{D^{-2}}(p_0, p_2) = 2\pi k$, the ellipsoid has to roll along the geodesic arc that develops as $\overline{p_0 p_2}$ in the affine tangent space, otherwise it has to roll along a broken geodesic that develops as the equal sides of any planar isosceles triangle with base $\overline{p_0 p_2}$ and such that the length of the equal sides is a multiple of 2π greater than $\mathrm{dist}_{D^{-2}}(p_0, p_2)$.

7 Conclusion

We have presented an alternative proof to the complete nonholonomy, or controllability property, of the system consisting of an n-dimensional ellipsoid rolling without twist and without slip over its affine tangent space at a point. The main ingredients in this construction consist in showing how to realize the forbidden motions of twist and slip by rolling without twisting and without slipping. This forms the main contents of Sect. 6.

Acknowledgements This work was developed while the first author was visiting the University of Coimbra. The work of the second author was supported by FCT project PTDC/EEA-CRO/122812/2010.

References

1. Biscolla, L.O.: Controlabilidade do Rolamento de uma Esfera Sobre uma Superfície de Revolução, Ph.D. thesis, Univerity of São Paolo (2005)
2. Grong, E.: Controllability of rolling without twisting or slipping in higher dimensions. SIAM J. Control Optim. **50**(4), 2462–2485 (2012). https://doi.org/10.1137/110829581
3. Helgason, S.: Differential Geometry, Lie Groups and Symmetric Spaces. Academic Press, London (1978)
4. Hüper, K., Krakowski, K.A., Silva Leite, F.: Rolling Maps in a Riemannian Framework. Mathematical Papers in Honour of Fátima Silva Leite. Textos de Matemática **43**, 15–30. Departamento de Matemática da Universidade de Coimbra, Portugal (2011)
5. Johnson, B.D.: The nonholonomy of the rolling sphere. Am. Math. Mon. **114**(6), 500–508 (2007)
6. Kleinsteuber, M., Hüper, K., Silva Leite, F.: Complete controllability of the N-sphere - a constructive proof. In: Proceedings of 3rd IFAC workshop on langrangian and hamiltonian methods for nonlinear control (LHMLC'06). Nagoya, Japan, 19–21. July 2006
7. Krakowski, K.A., Silva Leite, F.: An algorithm based on rolling to generate smooth interpolating curves on ellipsoids. Kybernetika **50**(4), 544–562 (2014)
8. Krakowski, K.A., Silva Leite, F.: Smooth interpolation on ellipsoids via rolling motions. In: Proceedings of PHYSCON 2013, San Luis Potosí, México, 26–29 August 2013
9. Krakowski, K.A., Silva Leite, F.: Why controllability of rolling may fail: a few illustrative examples. In Pré-Publicações do Departamento de Matemática, **22–26**, pp. 1–30. University of Coimbra (2012)
10. Sharpe, R.W.: Differential geometry: cartan's generalization of Klein's Erlangen program. Grad. Texts Math. **166**, (1997)

Methodological Approaches to Analyse Financial Exclusion from an Urban Perspective

Cristina Ruza-Paz-Curbera, Beatriz Fernández–Olit and Marta de la Cuesta-González

Abstract This paper aims to appraise the problem of financial exclusion in Spain after the process of banking system restructuration. The paper proposes a theoretical model for explaining the phenomenon of financial exclusion including both "access difficulties" and "difficulties in the use of financial services" as two dimensions that should be jointly considered. The main contribution of this paper is that it broadens the scope of financial exclusion from a theoretical and empirical point of view, and it also analyses the financial exclusion phenomenon at lower units of analysis that have not been previously explored: urban districts and municipalities. We considered Madrid and Barcelona as our scenarios of analysis. The methodological procedure was carried out in two steps: we first validate our theoretical model by applying canonical correlations, and secondly we carried out Quantile Regressions (QR) to estimate the different impact of financial exclusion's predictors at different points of the empirical distribution. The empirical results indicate a trend towards low-cost retail banking to serve the segment of less profitable customers and a pattern of branch disappearance more pronounced in those territories vulnerable from a socioeconomic point of view.

Keywords Financial exclusion · Consumer vulnerability · Financial crisis Geography and finance · Canonical correlation · Quantile regression · Inequality

C. Ruza-Paz-Curbera (✉) · M. de la Cuesta-González
Facultad CC. Económicas y Empresariales, UNED, Paseo Senda del Rey 11,
28040 Madrid, Spain
e-mail: cruza@cee.uned.es

B. Fernández–Olit
Faculty of Business and Communication, International University of La Rioja, UNIR,
c/ Almansa 101, 28040 Madrid, Spain
e-mail: beatriz.fernandez@unir.net

M. de la Cuesta-González
e-mail: mcuesta@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_20

1 Introduction

Access to commercial banking services is a basic condition for social integration in modern societies, and the quality of services provided by branches is still considered an essential indicator of financial inclusion. However, the concept of financial exclusion (FE) has evolved from the concept of financial exclusion as the lack of physical access to banking products and services, to a broader concept of exclusion that also considers the difficulties in the use of financial services.

Theories of geographical analysis initiated by Leyshon and Thrift [23] have focused on financial access mainly from a geographical point of view and exclusion was considered just a consequence of the lack or disappearance of bank branches in a territory.

The liberalization of the banking markets during the 80s entails a new orientation for the research of FE, considering also the impact of social and economic inequalities affecting communities. In addition, as the financial crisis has confirmed, consumer vulnerability in the financial market may involve a lack of control by the user (and abuse by the entity) leading to an unhealthy dependence, irrational decisions (Hill and Kozup [17]; De Meza et al. [8]) and to the general misuse of financial services. Thus, the assessment of financial exclusion should not only consider the quantity of financial services available, but also the use people make of the services provided.

On these grounds the paper evaluates the level of branch reduction and branch saturation as proxies of financial exclusion in terms of access difficulties and use difficulties, respectively. The main contribution of this paper is twofold. First, it broadens the scope of financial exclusion from a theoretical and empirical point of view, and secondly we focus on the financial exclusion phenomenon at lower units of analysis that will provide us with new insights into the determinants of the problem, that were not revealed by previous studies carried out at national or province levels. In particular, we focus the attention on the two biggest urban areas in Spain, whose banking network were severely reduced during the crisis.

The analysis of results will be especially relevant for policymakers interested in preventing financial exclusion problems as well as for the whole banking industry in designing new strategies in this scenario.

The paper is structured as follow: The first section describes the phenomenon of financial exclusion from the perspective of use difficulties associated with the consumption of financial services, and its relation with vulnerable consumption theory. The second section focuses on defining and empirically appraising a model of branch closure and branch saturation based on a set of social and economic variables for Madrid and Barcelona districts and municipalities. The third section presents the sample, methodology and empirical results. Finally, the paper ends with some conclusions and lines for future research.

2 Literature Review and Hypothesis Development

Leyshon and Thrift [24] defined financial exclusion as "those processes that serve to prevent certain social groups and individuals from gaining access to the financial system". Previous FE literature has mainly focused on financial access from a geographical point of view, and exclusion was considered a consequence of the lack of bank branches in a territory (Seaver and Fraser [26]; Evanoff [12]).

Kempson et al. [20] introduced the idea of FE as a product of diverse barriers (access, conditions, price, marketing and self-exclusion), which produce difficulties in accessing and using banking services (Devlin [9]; Anderloni and Carluccio [5]).

In this paper we adopt the term "difficulties of use" from Gloukoviezoff [15]. He defines FE as the process whereby people encounter such great difficulties to either access or use financial services that they can no longer lead a normal social life. FE then refers to "the mismatch between the way products are sold to customers or the characteristics of financial services and the needs of people". This has been especially important in the financial crisis, where both the lack of customer control and the mis-selling of financial products have led to an increase in over-indebtedness and, subsequently, financial and social exclusion (evictions, inequality and economic marginalisation). Thus, the assessment of financial exclusion should not only consider the quantity of financial services available, but also the use people do of the services provided.

Devlin et al. [10] attempted to measure "fairness in financial services" and found differences in customer care related to the type of customer considered.

People in a vulnerable situation need banking products and services with conditions adapted to their particular needs. They face serious risks when they misuse banking services, particularly elderly people, immigrants, and people who are unemployed or in a situation of working precariousness or poverty. Gloukoviezoff [15] emphasized their need for personal financial advice to avoid the pernicious effects of inadequate product selection. This advice has traditionally been supported by bank branches officers, which are crucial for those collectives. Nevertheless, the banking restructuration has reduced the network and increased the number of people attended in each branch, which potentially can lead to a reduction in the quality of services delivered. We wonder if this relocation of branches has generated an even more pronounced branch saturation in those areas with worse socioeconomic standards. Although online banking could reduce this effect, experiences combining financial and technologic literacy to promote the use of online banking among low-income individuals have not produce positive results (Servon and Kaestner [27]). Moreover, a recent EU survey showed that 72% of unbanked individuals (defined as those without a bank account) were not interested in banking online (Mori [25]). As Hogg et al. [18] stated, consumers who lack access to the information or technology required to take part in the new "virtual" markets, like the financial market, are considered the new kind of vulnerable consumer of the current knowledge-based economy.

Financial exclusion in terms of use difficulties is not a homogeneous problem and it cannot be solved by the simple presence of bank branches or, alternatively,

electronic banking. It seems necessary to analyse the problem focusing on the needs of the different groups of population, and particularly on those facing the highest risk of being financially excluded.

The most common factors of risk in terms of financial exclusion in developed countries could be defined by the socioeconomic characteristics of population: Low income households or individuals (Devlin [9]; Anderloni et al. [4]; Karp and Nash-Stacey [19]; FDIC [13]); immigrants or minorities (Anderloni and Carluccio [5]; Devlin [9]; Karp and Nash-Stacey [19]; FDIC [13]); age - mainly youths- (Anderloni et al. [4]; Karp and Nash-Stacey [19]; FDIC [13]); unemployment, particularly not belonging to the formal labor force (Anderloni and Carluccio [5]; Devlin [9]; Anderloni et al. [4]; Karp and Nash-Stacey [19]; FDIC [13]); and single parent house-

Table 1 Variables and information sources included in analysis focused on Spain

Study	Variable	Determinants	Method
Alamá et al. [2]	Number of branches per municipality (Spain)	Unemployment (proxy of income per capita)	Poisson regression model within the framework of a GLMM (Generalised linear mixed model)
		Population density	
		Foreign population	
		Province and municipality of origin of banking entity	
		Number of branches of other typology of banking entities	
Alamá and Tortosa-Ausina [3]	Number of branches per municipality (Spain)	Low income	Quantile regression based on the database of Anuario Economico La Caixa
		High unemployment	
		Social housing	
		General and retail commercial activities	
		Construction activities	
		Tourism	
		Population density	
Bernad et al. [6]	Number of branches in generally low-income and high-income municipalities (Spain)	Population	Ordinary Least Squares (OLS) based on the model of Lanzillotti and Saving [22]
		Population density	
		Income	

Source Own elaboration

holds (Anderloni et al. [4]; FDIC [13]). To a lesser extent, the low educational level (Devlin [9]; FDIC [13]) and the low level of financial literacy (Anderloni and Carluccio [5]) have also been studied among the determinants of FE. Table 1 resumes the socio-economic factors included in studies, focused on Spain, regarding the risk of financial exclusion due to irregular distribution of banking branches (at province or national levels).

The settlement of branches in vulnerable urban areas is no longer considered to be a duty to maintain the socio-economic equilibrium among neighbourhoods (Dysmki, [11]; Leyshon and Thrift [23]). Thus, the analysis of urban territories is proposed to identify discrimination in the quality of banking services attributed to social inequalities and the restructuration of the banking sector. This restructuration has reduced the diversity in the whole banking sector, with the nearly elimination of saving banks in countries like Spain. As a consequence, the number of branches has decreased and it have resulted in an "overloading" of remaining branches. The research model is based on Fig. 1.

Community and context based models integrate both individual and group factors, as well as potential external shocks. This seems to be an appropriate framework to analyse the complex nature of FE and underbanking processes which are phenomena simultaneously affected by static facts (e.g. deprived communities) and dynamic ones (e.g. diminishing of branch network, rise of unemployment and accelerated use of new technologies).

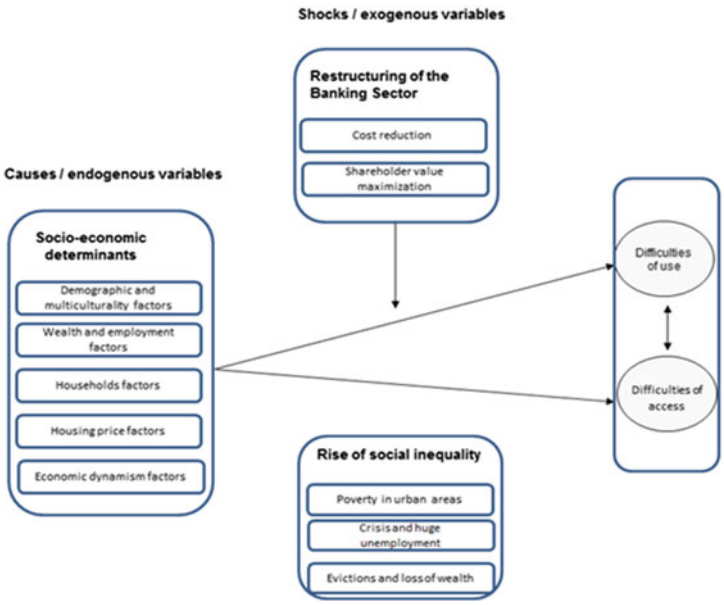


Fig. 1 Assessment model *Source* Own elaboration

We propose as hypotheses of analysis according to the model of Fig. 1 the following:

H1: Rates of branch closure (access difficulties) have been higher in territories with greater socio-economic vulnerability.

H2: The deterioration of the quality of banking services (use difficulties) has been greater in territories with greater socio-economic vulnerability.

3 Sample, Methodology and Results

The analysis has addressed the districts and municipalities belonging to the two largest Spanish Metropolitan Areas (MA), Madrid and Barcelona, and we use information of 63 municipalities and 31 districts. In Madrid, the covered territories include 27 municipalities of the MA (2,351,430 inhabitants in 2013) and the 21 districts of the city centre (3,215,633 inhabitants), following the definition of this MA stated by García and Sanz [14]. Regarding Barcelona, we include the 10 districts of Barcelona city (1,611,822 inhabitants) and the 36 municipalities of its MA (1,645,887 inhabitants), following the definition of Idescat[1] We define three statistical models. First, we use canonical correlations to appraise the correct definition of the underlying theoretical model regarding causes and consequences of recent FE processes in Spain. The second model assesses the determinants of 'branch reduction' (BRANCHREDUCT), while the third model focuses on the difficulties of use defining 'branch saturation'[2] (INHABBRANCH), as a proxy of the quality of financial services. Second and third models are analysed applying quantile regressions.

The dependent variables were obtained from the Historic Archive of the Banking Guide of Ediban[3] for the years 2008 and 2013. This database offers detailed information about all bank branches in Spain, allowing us to classify them by district and municipality using postal codes and addresses.

The independent variables included in the model were the socio-economic determinants previously identified in the literature (see Table 1).[4] Population density (POPDENS) is considered a classic determinant of market attractiveness for branch settlement, as well as multicultural population, represented by the rate of immigrants of the four main nationalities (IMMIGR). With regard to the labor and household economic variables, unemployment rate (UNEMPLOY) was selected as a proxy

[1] Idescat is the Statistics Institute of Catalonia Autonomous Community.

[2] The relation between the number of inhabitants and the number of bank branches has previously used in the study of FE in Spain by Alama et al. [2].

[3] www.maestre-ediban.com/.

[4] We have considered a broader number of factors included in Table 1 as independent variables. Nevertheless, after multicollinearity analysis we have selected the variables included in Table 2.

Table 2 Variables and information sources included in the analysis

Type of variable	Year	Source
Dependent variables		
Branch reduction	2008–2013	Historic Archive of the Banking Guide of Ediban
Inhabitants per branch	2013	Historic Archive of the Banking Guide of Ediban
Independent variables		
Socio-demographic indicators		
Population density	2013	Statistics Institutes of the Madrid and Catalonia Autonomous Communities
Percentage of population over 65	2013	National Statistics Institute. Continuous Municipal Register
Percentage of main immigrant nationalities	2013	National Statistics Institute. Continuous Municipal Register
Percentage of single-parent households *	2013	National Statistics Institute. Population and Housing Census
Socio-economic indicators		
Rate of unemployment per 100 inhabitants	2013	Own elaboration with data from the General Office of Statistics of Madrid and Barcelona City Councils, and Statistics Institutes of the Madrid and Catalonia Autonomous Communities (i)
Hotels per 1000 inhabitants	2013	Idem (i)
Variation in housing price *	2008–2013	Historic Database of Idealista Real Estate Agency. www.idealista.com
Rate of Internet access *	2011	National Statistics Institute. Population and Housing Census

The number of observations in all territories is 94 (48 MAD+46 BCN), except in those with *: SINGLEPAR - 90 (48 MAD+42 BCN); HOUSEPR - 82 (47 MAD+35 BCN) and INTERN 86 (48 MAD+38 BCN).
Source Own elaboration

of the profitability-risk combination (from the banking perspective), as well as the dynamism of tourism (TOURDYNAM), and the rate of variation in housing prices (HOUSEPR), considering that those territories more affected by depreciation have increased their social vulnerability. This study includes new variables in the study of financial exclusion in Spain, as they have become significant in other developed countries like the European Union and the USA (Anderloni et al. [4]; FDIC [13]): percentage of the population over 65 years of age (POP65), and rate of households with single-parent (SINGLEPAR), which help to better represent the demographic structure of territories.[5] We have included an innovative variable, in order to represent the recent changes in the banking channels: the rate of households with Internet

[5]These collectives have also been defined to face a higher risk of use difficulties with banking services (Anderloni and Carluccio [5]; Gloukoviezoff [15]; Devlin [9]).

Table 3 Descriptive statistics of the variables

	N observations	Mean	Standard deviation	Median	Q25	Q75
Dependent variables						
Branch reduction (%)	94 (100%)	24.40	0.11	26.28	17.94	31.24
Inhabitants per branch	94 (100%)	1477.56	435.79	1440.86	1239.80	1675.83
Independent variables						
Population density	94 (100%)	7561.87	8764.85	3705.44	1118.36	12255.35
Percentage of population over 65 (%)	94 (100%)	15.22				
0.05	14.89	11.68	18.34			
Percentage of main immigrant nationalities (%)	94 (100%)	4.91	0.04	4.02	2.24	6.72
Percentage of single-parent households (%)	90 (95.7%)	5.03	0.01	5.04	4.48	5.77
Rate of unemployment per 100 inhabitants	94 (100%)	8.09	2.24	7.97	6.12	9.78
Hotels per 1000 inhabitants	94 (100%)	0.13	0.28	0.06	0.01	0.13
Variation in housing price (devaluation, %)	82 (87.2%)	26.13	0.07	31.14	25.74	35.80
Rate of internet access (%)	86 (91.5%)	62.88	0.08	66.53	62.82	71.00

Source Own elaboration

access (INTERN) is a relevant factor to explore alternative services such as online banking. Table 3 summarizes the descriptive statistics of dependent and independent variables.

Canonical Correlations

We can define canonical correlation analysis as a multivariate technique to identify and measure the strength of association between two sets of variables; one can be interpreted as the causes of a phenomenon and the other as the consequences. It is particularly adequate when we are dealing with high number of variables that are

grouped into two sets: one is composed by exogenous variables or predictors and the second group is formed by the endogenous variables to be studied. This technique is particularly suitable if it is observed the presence of high correlation among selected variables.

Applying canonical correlation analysis is twofold. First, it permits to appraise if the underlying theoretical model has been correctly defined, as long as it measures the strength between a group of causes and consequences for a certain phenomenon, when both groups are considered multidimensional. Secondly, if we have to reduce the number of predictors of our selected model due to multicollinearity we will be able to determine which variables are contributing to a higher extent to the variability of the exogenous or endogenous variables on the whole. The use of canonical correlations is innovative in the study of FE and reinforces its multi-causal character. The analysis is focused on latent variables called canonical variates (which are not directly observed) constructed in such a way that best explain the variability within the two sets of variables (which are directly observed, i.e. our selected variables).

These canonical variates are sequentially constructed in pairs, so at each stage of the procedure we will have a canonical variate for the set of explained variables (set 1) and another variate for the set of predictors (set 2).

Let define for multiple x and y, two canonical variates: CV_{X1} and CV_{Y1}

$$CV_{X1} = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n \tag{1}$$

$$CV_{Y1} = b_1y_1 + b_2y_2 + b_3y_3 + \ldots + b_my_n \tag{2}$$

The process consists of estimating the canonical weights ($a_1 \ldots a_n$) for the first set and ($b_1 \ldots b_n$) for the second set, so that they maximise:

$$MaxCorrelation(CV_X, CV_Y) \tag{3}$$

And then repeat the process until we identify:

$$CV_{xm} = a_{m1}x_{m1} + a_{m2}x_{m2} + a_{m3}x_{m3} + \ldots + a_{mn}x_n \tag{4}$$

$$CV_{ym} = b_{m1}y_{m1} + b_{m2}y_{m2} + b_{m3}y_{m3} + \ldots + b_{mn}y_n \tag{5}$$

that jointly verifies:

$$C_m \; correlation \; is \; maximum \tag{6}$$

$$CV_{ym} = b_{m1}y_{m1} + b_{m2}y_{m2} + b_{m3}y_{m3} + \ldots + b_{mn}y_n \tag{7}$$

$$Cor(CV_{xj}CV_{xk}) = 0 \qquad \forall j \neq k \tag{8}$$

$$Cor(CV_{yj}CV_{yk}) = 0 \qquad \forall j \neq k \tag{9}$$

Table 4 Canonical correlation analysis

Variables	Standardised coefficients	
	Root 1	Root 2
Set 1 (Problem)		
Branch reduction (BRANCHRED)	0.3516	0.9992
Inhabitants per branch (INHABITBR)	1.0592	−0.0025
Set 2 (Predictors)		
Population over 65 (POP65)	0.4068	1.2589
Population density (POPDENS)	−0.0383	−0.0852
Immigrants (IMMIGR)	0.5635	−0.9023
Unemployment (UNEMPLOY)	−0.0783	0.3397
Single-parent households (SINGLEPAR)	0.5132	0.1073
Price of housing (HOUSEPRICE)	−0.2673	−0.0384
Internet access (INTERN)	−0.2385	−0.0699
Touristic dynamism (TOURDYNAM)	−0.0072	0.9268
Canonical correlation	0.9228	0.7518

Source Own elaboration

Table 5 Test of overall significance

	Statistic	F	Prob > F
Wilkslambda	0.0645	7.7052	0.0000
Pillais trace	1.4166	6.6789	0.0000
Lawley–Hotelling trace	7.0336	8.7921	0.0000
Roys largest root	5.7338	15.768	0.0000

Source Own elaboration

The canonical root is defined as the smallest number of variables in each of the two sets. In our case the canonical root or dimension is equal to the number of variables in the smallest set: two. In Table 4 we show the standardized coefficients.

The standardized canonical coefficients are interpreted in a manner analogous to interpreting standardised regression coefficients. Under the first specification or root 1 we see that the correlation between the set of causes and consequences of the problem of financial exclusion is above 92%. Then we can argue that socioeconomic characteristics of a territory are highly correlated with branching behavior. This result gives support to the theoretical model under which we stated that FE problem should be appraised from a broad dimension that simultaneously include physical access to banking services (through the branch network) and also a qualitative access based on customised personal support.

We test the general fit of the model according to Wilk's, Pillai's, Lawley–Hotelling and Roy's multivariate criteria and we find that all of these tests are significant at 5% (Table 5).

Table 6 Test of coefficients significance

	Coefficient	St. Error	t	P > ItI	h^2 (%)
Root 1					
Branch reduction	6.368	1.5141	4.21	0.000	1.0000
Inhabitant per branch	0.0018	0.0001	12.67	0.000	1.0000
Population over 65	0.0427	0.0163	2.61	0.000	0.7521
Population density	0.000	0.0001	−0.26	0.000	0.2099
Immigrants	20.6165	11.9674	1.72	0.000	0.5418
Unemployment	−0.0425	0.1861	−0.23	0.000	0.7213
Single-parent households	69.8331	21.0973	3.31	0.000	0.1478
Price of housing	−3.5657	2.5019	−1.43	0.000	0.5731
Internet access	−4.2555	4.5892	−0.93	0.000	0.4377
Touristic dynamism	−0.019	0.5307	−0.04	0.000	0.0656

Source Own elaboration

Because Wilks's λ represents the variance unexplained by the model, the factor 1− λ yields the full model effect size in an r^2 metric. In our model the whole r^2 type effect size was 0.94. Following, we want to identify what raw coefficient for each of the canonical variates is individually significant. For the first dimension (root 1) "branch reduction" and "inhabitants per branch" share some variability with one another, and are statistically significant. Within the second set of predictors, "immigrants" and "single parent households" are statistically significant. The second dimension is not significant and no attention will be paid to its coefficients or to the Wald tests.

The last column of Table 6 shows the communality coefficient h2 (%) as the proportion of variance in each variable that is explained by the canonical function that is interpreted. This statistic indicates how useful each observed variable was for the entire analysis. The "causes" variables with a high explanatory power are "population over 65" (h2 = 75.21%), "unemployment" (h2 = 72.13%), "price of housing" (h2 = 57.31%) and "immigrants" (h2 = 54.18%), while consequences of FE (access difficulties and use difficulties) are both fully relevant.

Once the theoretical specification of our model has been validated by canonical correlation analysis, the next step is to perform two separated regressions of the variables under study: "branch reduction" (access exclusion) and "inhabitants per branch" (use difficulties).

Quantile Regressions

The two models to be estimated are:

Model 1:

$$(\text{BRANCHREDUCT}) = \beta_1(\text{POPDENS}) + \beta_2(\text{POP65}) + \beta_3(\text{IMMIGR}) + \beta_4(\text{UNEMPLOY}) + \beta_5(\text{SINGLEPAR}) + \beta_6(\text{HOUSEPRICE}) + \beta_7(\text{INTERN}) \beta_8(\text{TOURDYNAM})$$

Model 2:
(INHABBRANCH)= β_1(POPDENS) + β_2(POP65) + β_3(IMMIGR) + β_4(UNEMPLOY) + β_5(SINGLEPAR) + β_6(HOUSEPRICE) + β_7(INTERN) β_8(TOURDYNAM)

Following Alamá and Tortosa-Ausina [3] we apply quantile regressions (QR) to estimate conditional quantile functions of the response variable given a dependent variable defined as a linear function of the covariates. The underlying assumption of QR is that impacts on the response variable should not be the same over the entire conditional distribution (Koenker [21]).

QR has considerable appeal for various reasons: it is robust to the presence of outliers; it estimates the impact of covariates on location (central and noncentral) and scale parameters; it does not impose restrictions on the error term like OLS and fits nonnormal and heteroscedastic data. Finally, QR also takes into account how changes in the covariates might affect the underlying shape of the distribution of the response variable (Hao and Naiman [16]), which is of great interest in our particular case because the "financial exclusion" phenomenon is itself an extreme case of study.

Table 7 Determinants of branch reduction using quantile regression

	OLS	QR$\vartheta = 5$	QR$\vartheta = 50$	QR$\vartheta = 80$	QR$\vartheta = 99$
Constant	0.2905**	0.005*	0.3474*	0.9155	0.9215**
POP65	−0.0055	−0.0075**	−0.0058	−0.0061	−0.0063**
	(0.0014)	(2.112 e-17)	(0.0039)	(0.0052)	(2.99 e-19)
POPDENS	3.625e-07	−3.17 e-05**	1.53 e-05	2.74 e-06	7.864 e-07**
	(1.29 e-06)	(1.978 e-20)	(3.332 e-06)	(3.98 e-06)	(2.787 e-22)
IMMIGR	1.3676	1.1929**	2.1166	4.2813**	2.3952**
	(1.0432)	(1.316 e-14)	(2.75)	(2.0846)	(2.425 e-16)
UNEMPLOY	−0.0076	0.0065**	−0.0165	−0.0577*	−0.0326**
	(0.0162)	(1.895 e-16)	(0.0394)	(0.0341)	(3.901 e-18)
SINGLEPAR	−0.6147	−7.2374**	1.2556	2.6636	−0.1291**
	(1.8391)	(2.57 e-13)	(4.5024)	(6.8099)	(3.586 e-16)
HOUSEPRICE	0.0216	−0.1735**	0.0478	0.1804	0.1127**
	(0.2181)	(2.895 e-15)	(0.5782)	(0.5087)	(4.69 e-17)
INTERN	0.0523	0.7923**	−0.1234	−0.7047	−0.6272**
	(0.4)	(5.445 e-15)	(1.033)	(0.9977)	(8.718 e-17)
TOURDYNAM	−0.1013	−0.0868**	−0.1265	−0.1871*	−0.1701**
	(0.0462)	(6.338 e-16)	(0.1087)	(0.1011)	(8.892 e-18)
R^2	56.52	53.53	38.7	50.45	63.68

Source Own elaboration

In Table 7, coefficients appear more significant at the more extreme quantiles (Q = 5 and 99%). The standard errors are higher for median regression (QR = 50%), reflecting higher precision of estimation at the two extremes of the distribution. One reason for coefficient differing across quantiles is the presence of heteroskedastic errors. We run the Breusch–Pagan test and we could not reject the null hypothesis of homoscedasticity. Moving now to quantile regressions at different points of the distribution, we should precise that the higher quantile (QR99) corresponds to those territories with higher branch abandonment. In those territories the variables that contribute to a higher extent to bank's closure of branches is the presence of the immigrant population (+2.39) and house price devaluation (+0.11). The rest of variables show negative coefficients.

Next, we will analyse results for the second model based on the use difficulties approach (Table 8).

Some conclusions can be drawn from the empirical results. The main determinant of branch saturation is the immigrant population that lives in a district or municipality and the rate of single parents for QR 99. However, when we move to the lower tail of the distribution we observe that higher level of saturation corresponds to a higher presence of population over 65 years, higher unemployment and higher single parent households. In addition, we can argue that saturation is more pronounced in the less dynamic and prosperous territories where young people tend to settle, where house

Table 8 Determinants of branch saturation using quantile regression

	OLS	QR$\vartheta = 5$	QR$\vartheta = 50$	QR$\vartheta = 80$	QR$\vartheta = 99$
Constant	302.6147*	−651.7435**	871.63	−819.7348	428.0389**
	(2387.0301)	(0.00006)	(6481.656)	(2709.15)	(8.17 e-13)
POP65	2.4649	17.6248**	−0.2186	1.9825	−6.6732**
	(9.7471)	(2.9 e-07)	(28.419)	(13.8452)	(3.24 e-15)
POPDENS	−0.0007	−0.02501**	0.0015	0.0091	0.0091**
	(0.0088)	(3.03 e-10)	(0.0239)	(0.0121)	(2.90 e-18)
IMMIGR	15343.686**	−11169.24**	22114.93	25996.12**	25923.28**
	(7125.6771)	(0.0002)	(18002.74)	(7659.39)	(2.45 e-12)
UNEMPLOY	−48.6741	265.3872**	−182.2445	−221.427**	−206.02**
	(110.8163)	(3.09 e-06)	(258.9032).	(109.1023)	(4.08 e-14)
SINGLEPAR	33485*	22.0513**	37876.46	53998.84**	52702.23**
	(12561.834)	(0.0004)	(30205.44)	(12737.37)	(3.89 e-12
HOUSEPRICE	−1744.0273	−2788.339**	−1900.244	−2702. 81*	−1544.488**
	(1489.7115)	(0.0004)	(4166.939).	(2585.45)	(5.38 e-13)
INTERN	−1988.2449	−685.0755**	−2105.529	−608.0628	−2004.248**
	(2732.5591)	(0.0007)	(7202.686)	(347.02)	(9.16 e-13)
TOURDYNAM	−366.7635*	668.8571**	−565.6039	−0.1423*	−815.9901**
	(316.002)	(9.54 e-06)	(818.7199)	(5.47 e-16)	(9.63 e-14)
R^2	82.00	60.22	61.87	70.22	76.54

Source Own elaboration

prices variation tend to be lower, where houses have access to internet to a lesser extent and finally where the economic dynamism is relatively lower compared to other areas.

Our results partly agree with the conclusions of previous studies carried out in Spain (Alamá et al. [2] immigrant population- ; Alamá and Tortosa-Ausina [3] low income and economic activity- Bernad et al. [6] low income) as well as in other European urban areas (Aalbers [1] ethnic diversity, low income and depreciation of housing-). However most of these studies have widely applied OLS techniques, which masks the different reality revealed at the tails of the empirical distribution. As we have argued, predictors of financial exclusion are not the same when focusing the attention on the areas with lower and higher rates of branch disappearance, respectively, or branch saturation. By analysing these areas separately (applying quantile regressions) we will be better able to capture how variables impact the dependant variable in each case. This is one of the main contributions of the paper in the financial exclusion arena.

4 Conclusions

From our assessment theoretical model we can highlight the main empirical findings of this research as follows.

Empirical results at Madrid and Barcelona city level reveal that our first hypothesis H1 (There have been higher rates of branch closure -access difficulties- in those territories with higher socio-economic vulnerability) holds when considering immigrant population and variation of housing prices.

At the city areas of Madrid and Barcelona our empirical results support H2 (There have been a higher quality deterioration of bank services -use difficulties- in those territories with higher socio-economic vulnerability) when considering immigrants and single parent households.

To conclude, debranching is creating inequalities in the distribution of resources that reinforces the structural vulnerability of some areas. In highly bancarised countries like Spain FE is not only a problem of physical distance to the branch (Carbó and Rodríguez [7]), but also a matter of growing inequality in the use of banking services, our empirical results have demonstrated that causes and consequences of FE are closely interrelated as we have analysed by applying canonical correlations. Based on our results of quantile regressions, we can argue that the unequal pattern of disappearance of branches in the two main Spanish cities of Madrid and Barcelona reflects a banking reallocation of resources highly depending on the socioeconomic

characteristics of the territories, revealing a trend towards "low-cost" retail banking to serve the segment of less profitable customers and a pattern of branch disappearance more pronounced in socioeconomic vulnerable territories.

Our empirical results confirm that low-cost business practices could be reaching the banking sector in Spain. The immediate consequence is the aggravation of financial discrimination in the urban areas most severely damaged by the crisis. These results are especially relevant for policymakers interested in preventing financial exclusion problems and the misuse of financial products. It can be also interesting for the whole Spanish banking industry, to alert them about the increase of less formal financial providers as a feasible alternative for those excluded from traditional banking.

Acknowledgements The authors acknowledge the financial support of Fundación de las Cajas de Ahorro (FUNCAS).

References

1. Aalbers, M.B.: What types of neighbourhoods are redlined? J. Hous. Built Environ. **22**(2), 177–198 (2007)
2. Alamá, L., Conesa, D., Forte, A., Tortosa–Ausina E.: The geography of Spanish bank branches. J. Appl. Stat. **42**(4), 722–744 (2015)
3. Alamá, L., Tortosa–Ausina E.: Bank branch geographic location patterns in Spain: some implications for financial exclusion. Growth Change. **43**(3), 505–543 (2012)
4. Anderloni, L., Bayot, B., Bledowsk, P., Iwanicz-Drozdowsk, M., Kempson, E.: Financial Services Provision and Prevention of Financial Exclusion. European Commission, Brussels (2008)
5. Anderloni, L., Carluccio, E.M.: Access to bank accounts and payment services. New Frontiers in Banking Services, pp. 5–105. Springer, Berlin (2007)
6. Bernad, C., Fuentelsaz, L., Gomez, J.: Deregulation and its long-run effects on the availability of banking services in low-income communities. Environ. Plan. A **40**(7), 1681 (2008)
7. Carbó, S., Rodríguez, F.: Concepto y evolución de la exclusión financiera: Una revisión. Cuadernos de Información Económica **244**, 73–83 (2015)
8. De Meza, D., Irlenbusch, B., Reyniers, D.: Financial capability: a behavioural economics perspective. Financial Services Authority (2008)
9. Devlin, J.F.: A detailed study of financial exclusion in the UK. J. Consum. Policy **28**(1), 75–108 (2005)
10. Devlin, J.F., Sanjit, K.R., Sekhon, H.: Perceptions of fair treatment in financial services. Eur. J. Mark. **48**(7), 1315–1332 (2014)
11. Dymski, G.A.: Immigration, finance, and urban evolution: an illustrative model with Los Angeles case study. Rev. Black Polit. Econ. **30**(4), 27–50 (2003)
12. Evanoff, D.: Branch banking and service accessibility. Money Credit Bank. **20**(2), 191–202 (1988)
13. FDIC.: National Survey of unbanked and underbanked households. Federal Deposit Insurance Corporation (2013)
14. García A., Sanz B.: Atlas de la Comunidad de Madrid en el umbral del siglo XXI. Comunidad de Madrid and Universidad Complutense de Madrid (2002)
15. Gloukoviezoff, G.: From financial exclusion to overindebtedness: the paradox of difficulties for people on low incomes? New Frontiers in Banking Services, pp. 213–245. Springer, Berlin (2007)

16. Hao, L., Naiman, D.Q.: Quantile Regression, Quantitative Applications in the Social Sciences. Sage, Thousand Oaks (2007)
17. Hill, R.P., Kozup, J.C.: Consumers experiences with predatory lending practices. J. Consum. Aff. **41**(1), 29–46 (2007)
18. Hogg, M.K., Howells, G., Milman, D.: Consumers in the knowledge-based economy (KBE): what creates and/or constitutes consumer vulnerability in the KBE? J. Consum. Policy **30**(2), 151–158 (2007)
19. Karp, N., Nash-Stacey, B.: Technology, Opportunity and Access: Understanding Financial Inclusion in the US. BBVA Research paper, No. 15/25 (2015)
20. Kempson, E., Whyley, C., Caskey, J., Collard, S.: In or out?: financial exclusion: literature and research review. Financial Services Authority, UK (2000)
21. Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)
22. Lanzillotti, R.F., Saving, T.R.: State branching restrictions and the availability of branching services. J.Money Credit Bank. **1**, 778–788 (1969)
23. Leyshon, A., Thrift, N.: Geographies of financial exclusion: financial abandonment in Britain and the United States. Trans. Inst. Br. Geogr. **20**(3), 312–341 (1995a)
24. Leyshon, A., Thrift, N.: Financial exclusion and the shifting boundaries of the financial system. Environ. Plan. A **28**(7), 1150–1156 (1995b)
25. Mori, I.: Road to Inclusion. A look at the financially underserved and excluded across Europe, MasterCard and Ipsos MORI (2013)
26. Seaver, W.L., Fraser, D.R.: Branch banking and the availability of banking services in metropolitan areas. J. Financ. Quant. Anal. **14**(01), 153–160 (1979)
27. Servon, L.J., Kaestner, R.: Consumer financial literacy and the impact of online banking on the financial behavior of Lower Income bank customers. J. Consum. Aff. **42**(2), 271–305 (2008)

Prospective Study About the Influence of Human Mobility in Dengue Transmission in the State of Rio de Janeiro

Bruna C. dos Santos, Larissa M. Sartori, Claudia Peixoto, Joyce S. Bevilacqua and Sergio M. Oliva

Abstract Dengue is a human arboviral disease transmitted by Aedes mosquitoes and it is currently a major public health problem in which around 2–5 billion people are at risk of infection each year. Climate changes and human mobility contribute to increase the number of cases and to spread the disease all around the world. In this work, the influence of human mobility is evaluated by analyzing a sequence of correlations of dengue incidence between cities in southeastern Brazil. The methodology initially identifies the cities were the epidemy begins, considered as *focus* for that epidemic year. The strength of the linear association between all pairs of cities were calculated identifying the cities which have high correlations with the focus-cities. The correlations are also calculated between all pairs considering a time lag of 1, 2 or 3 weeks ahead for all cities except the focus ones. Centred differences of the notification number are used to detect the outbreaks. The tests were made with DATASUS-SINAN data of the state of Rio de Janeiro, from January 2008 to December 2013. Preliminary results indicate that the spread of dengue from one city to another can be characterized by the development of the sequence of shifted correlations. The proposal may be useful to consider control strategies against disease transmission.

Keywords Dengue · Human mobility · Correlation · Rio de Janeiro

B. C. dos Santos (✉) · L. M. Sartori · C. Peixoto · J. S. Bevilacqua · S. M. Oliva
Instituto de Matemática e Estatística, Universidade de São Paulo,
Rua do Matão, 1010, São Paulo, Brazil
e-mail: brunacs@ime.usp.br

L. M. Sartori
e-mail: larissa@ime.usp.br

C. Peixoto
e-mail: claudiap@ime.usp.br

J. S. Bevilacqua
e-mail: joyce@ime.usp.br

S. M. Oliva
e-mail: smo@ime.usp.br

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_21

1 Introduction

Dengue virus infects each year about 300 million people worldwide and nearly 90 million of them develop the classic symptoms of the disease, such as fever, headache and nausea. Currently, dengue is endemic in more than 100 countries in Africa, America, Asia and Oceania [9]. In Brazil, the first documented occurrence was in Roraima, 1981–1982, and the first huge epidemic was in Rio de Janeiro city, 1986 [17]. The largest outbreak in Brazil occurred in 2013, accounting for around 1.5 million of notified cases [13]. Dengue is transmitted primarily by *Aedes* mosquitoes, particularly *Aedes aegypti*. The disease manifests in tropical and subtropical areas, which climatic conditions favor the development of eggs into larvae and mosquitoes. In Brazil circulate four strains of the virus, known as DEN-1, DEN-2, DEN-3 and DEN-4 of the family Flaviviridae, *genus Flavivirus* [3].

Factors such as population growth, global warming, rural-urban migration, environmental deterioration and the quality of basic sanitation, are some of the causes for the increase in infectious disease transmitted by vectors [12, 22]. Although there are not a consensus about disease's persistence, recent studies suggested the human mobility may be responsable for the emergence and reemergence of some diseases, in both the direct and indirectly transmitted [1, 17, 21]. Chikungunya and Zika outbreaks in Brazil are examples of diseases that have emerged in the country lately, and until recently, Chikungunya had only been detected in Africa, East Asia and India [10, 15].

Adams and Kappan [1] indicate that the spread of influenza and SARS (Severe Acute Respiratory Syndrome), from national to continental scale has been supported by the growth of airline transport network. In both global and local scales exist a daily traffic of people who move to work, tourism, etc. In the case of dengue, many people are asymptomatic then this scenario may be even more pronounced, because people may be spreading the disease to other places without even known they are infected [1, 4, 21]. The study developed in [1] highlights the role of human movement for the disease's persistence by establishing a dynamic on a hypothetical network. The authors observed that the understanding of human mobility can be used to map risk areas and provide targets for intervention and prevention. Stoddard et al. [21] investigated the relevance of human movement associated with vector behavior and how these two factors can increase the risk of exposure to disease due to human movement.

In general, most people have the same habit of daily mobility, in this work we analyze if the spread from one city to another can be explained by human mobility. Correlations between all pairs of cities were calculated considering that the beginning of the disease in each pair may be synchronized or not. The methodology is applied to a region composed by the municipalities of Rio de Janeiro State and all the border cities of Sao Paulo, Minas Gerais e Espirito Santo.

2 Materials and Methods

The state of Rio de Janeiro is located in southeastern Brazil, has a total population of 16.627.880 inhabitants [16]. About 96% of its population live in urban areas. The climate is tropical with an average temperature of 25 Celsius degrees throughout the year. Rio de Janeiro is one of the most visited place in Brazil, receiving tourists from all over the world during all seasons. These conditions, along with climate changes and increasing urbanization, ensure the mosquitoes proliferation and the disease maintenance [9].

Our analysis is based on data obtained from database of Notifiable Diseases Information System (Sistema de Informação de Agravos de Notificação - SINAN) an entity of Federal Government. In our study, the data considered were the weekly cases of dengue incidence from 2008 to 2013 for all cities of Rio de Janeiro and also the surrounding cities, totaling 130 cities [19]. The raw data were normalized by the urban population of each city [16].

The incidence considered significant was based on epidemiological alert thresholds defined by the Ministry of Health, therefore, there were excluded from the study, cities with incidence below than 300 cases per 100.000 inhabitants [13].

Defining a period of 52 weeks the methodology initially identify the cities that had outbreak, for instance, the cities which the number of notifications is equal or greater than 300 cases per 100 thousand inhabitants. Among this subset of cities a second cut is done excluding the cities which the total size of the population is less than 50 thousand inhabitants. After these two filters we selected the cities that first reach the incidence of 300 cases and define them as focus of the infection. Centred differences of the notification numbers were used to detect the outbreaks.

The correlations between all pairs of cities were calculated for the whole region with Pearson coefficient for two cases: Case 1: for all cities the period of analysis is defined from week 1 to week 52; Case 2: except by the focus, the period for the other cities is defined with a delay of 1, 2 or 3 weeks. A high correlation with delays between two cities (C_j, C_k), $j, k = 1, \cdots, m$, where m is the total number of cities being analyzed, suggests that the outbreaks have a time lag of n-weeks, which may indicate that the disease migrates from city C_j to city C_k. We define that the correlation is significant if its value is greater than 0.8 and the significance $p - value$ is less than 0.05. Our hypothesis is that dengue spread from one city to another and it could be verified by the evolution of the sequence of n time lag calculated correlations.

We aimed to test essentially if the disease initiates at the same time all over the state. We were inspired by the work of Saba et al. [18] that used the correlation between the occurrences of cases of dengue between cities in the state of Bahia to build a network of mobility [11, 14]. The authors considered that the existence of correlation between cases of dengue in two cities corresponds to an edge of the graph. It is possible to see through the graphs of incidence that there is a time lag between the epidemic curves, then we considered reasonable to verify the hypothesis of human mobility in the spread of the disease through the development of correlations.

3 Results

The year 2008 was chosen for presentation of results due to the high incidence of reported cases and by present a well-defined qualitative behavior compared to other years, however, the whole period from 2008 to 2013 was analyzed. We defined the epidemiological year considering the period from January to December, because we are assuming that the disease is the same in the whole state.

Considering the two filters described by the methodology, among the 130 municipalities analyzed, 18 of these are part of the metropolitan area of Rio de Janeiro (MARJ), 3 of these are part of Baixada Litorânea, 3 of these are part of the Médio Parnaíba, 2 of these are part of the northwest region, 1 of northern region and one of Região da Costa Verde. From the incidence data observed for this year, the cities Angra dos Reis, Campos dos Goyatacazes, Niterói, Nova Iguaçu, Rio de Janeiro and Seropédica, were chosen as focus of the disease because they were the first cities to achieve 300 cases per 100.000 inhabitants.

Table 1 shows in the first column the pairs of cities with high correlation presented without delay, the second and third columns are the correlations with n-weeks of delay $n = 1, 2$. Correlation nn', $n' >= n$, $n, n' = 0, 1, 2, 3$, means that the correlation is evaluated between the time series of two cities shifted, respectively, by n and by n' weeks from the first week of the period of one year. If $n = 0$ and $n' > n$, the focus cities were fixed in the first position with no delay, and the other cities were shifted by n-weeks, giving Correlation 00, 01, 02 and 03. Intermediate correlations as Correlation 12, 13, 23 were calculated in order to try to explain some cases of high-type correlations between cities that are geographically distant.

Table 1 Pairs of Cities in the Rio de Janeiro State with correlation above 0.8. Correlation 00 means correlation between the cites C_j and C_k without delay. Correlation 01 is the correlation between the cities C_j and C_k with the city C_k shifted by one week. Correlation 12 means the city C_j shifted by one week correlated with the city C_k shifted two weeks

	Correlation 00	Correlation 01	Correlation 12
1	Nova Iguaçu - Niterói	Niterói - Itaboraí	Itaboraí - Cachoeiras de Macacu
2	Seropédica - Duque de Caxias	Duque de Caxias - Itaboraí	Itaboraí - Cachoeiras de Macacu
3	Rio de Janeiro - Duque de Caxias	Duque de Caxias - Magé	Magé - Rio Bonito
4	Niterói - Araruama	Araruama - Saquarema	
5	Niterói - Duque de Caxias	Duque de Caxias - Magé	Magé - São Pedro da Aldeia
6	Niterói - Rio de Janeiro	Rio de Janeiro - Magé	Magé - Rio Bonito
7	Campos dos Goyatacazes - Araruama	Araruama - Saquarema	
8	Cantagalo - Cordeiro	Cordeiro - Sto Ant. de Padua	Sto Ant. de Pádua - Porciúncula

From Table 1 is possible to observe that dengue begin simultaneously in the pairs presented in the first column of the Table 1, because high correlation was found with no delay. On the other hand, high correlations between nearby and distant cities with delay of 1 or 2 weeks were found (second and third columns of the Table 1), when one of the city of the pair is a focus.

Examples in which appear high correlation between dengue time series with relatively distant cities, could be an indicative of the role of human mobility in spreading the disease. According to Farias [6] is possible to highlight two types of commuting in the state of Rio de Janeiro: daily flows of short distance and greater frequency, mainly associated with trade and manufacturing industries (intra-regional level); and not daily flow of great distance and low frequency, associated with mining and construction industry inter-regional level.

In fact, such commutings may explain some of the correlations. The state has significant flow rates primarily concentrated in the metropolitan area of Rio de Janeiro, MARJ for short [7], explaining the high correlations independent of the existence of the delay between the cities of MARJ region, respectively, lines 1, 2, 3 and 6 of the Table 1.

On the other hand, intercensual analysis from 2000 to 2010 indicate a decentralization of the pendulum movement inside MARJ. A significant growth of the pendularyity outside MARJ was observed concentrated mainly between the northern regions and the coast. During the first decade of this century some urban centers in the state, especially Macaé, expanded its area of influence to the northern region, in particular to Itaperuna and to Baixada Litorânea. This movement could be observed in the pairs presented in the lines 4 and 7 of the Table 1.

The correlation between Magé and São Pedro da Aldeia in the line 5, may not necessarily be explained by human mobility, since there are few signs of mobility between these two cities. In especial, we see a great migration from São Pedro da Aldeia to Macaé and other cities that make up the region of OMPETRO [5]. As it is generally known, dengue is influenced by several factors such as climate, temperature, basic sanitation or public health policy, and in these cases it is not ruled out the hypothesis that the epidemic curves for these two cities have obtained correlation because the events may have occurred simultaneously but in an isolated manner. For these cases, we address that mobility is not responsible for high correlation.

In addition, for a more complete analysis, we also calculated the correlations including those cities which had urban population smaller than 50000 and greater than 10000, and that also reached more than 300 cases per 100 thousand inhabitants in 2008. About 66 cities were selected, the focus was Cantagalo that reached more than 300 cases in the fifth epidemiological week.

Cantagalo is characterized as an independent pole and correlates with other cities of the mountainous region of north and northwest parts of Rio de Janeiro State [2]. It has obtained correlation between Porciúncula and Santo Antônio de Padua with one and two weeks delay showed in Table 1, line 8. Although the 2010 census data indicates that exists considerable migration between metropolitan cities and the mountainous northwestern region of Rio de Janeiro, we did not find sufficient evidence to suggest that human mobility has been responsible for this association.

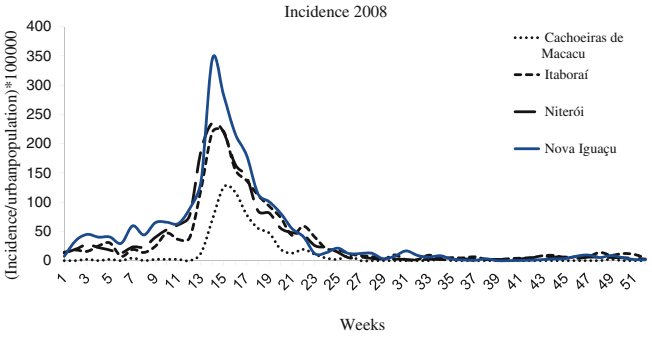


Fig. 1 Incidences of dengue cases in 2008 for cities that have correlation with Nova Iguaçu

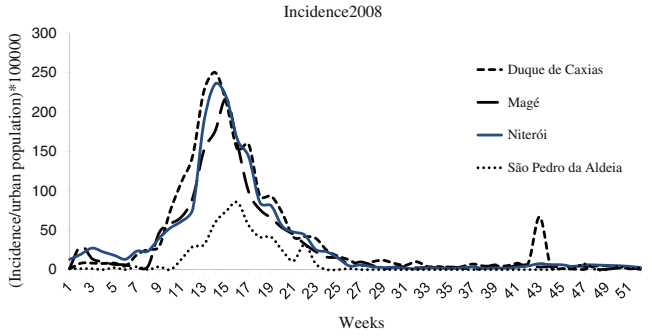


Fig. 2 Incidences of dengue cases in 2008 for cities that have correlation with Niterói

Figures 1, 2, 3 and 4 show the incidence data of DATASUS-SINAN observed in 2008. In Fig. 1 is presented the correlations obtained in line 1 of the Table 1. Nova Iguaçu was chosen as focus because this city was the first one to achieve the 300 cases (tenth epidemiological week). Nova Iguaçu and Niterói have high correlation with no shift (line 1, column 2); Itaboraí has higher correlation with Niterói with a shift of 1 week and finally Cachoeiras de Macacu has higher correlation with Itaboraí with a shift of 2 weeks.

In Fig. 2 are presented the correlations obtained in line 5 of Table 1. Niterói was chosen as focus because this city was the first one to achieve the 300 cases (tenth epidemiological week). Niterói and Duque de Caxias have high correlation with no shift (line 5, column 2); Magé has higher correlation with Duque de Caxias with a shift of 1 week and finally São Pedro da Aldeia has higher correlation with Magé with a shift of 2 weeks.

In Fig. 3 are presented the correlations obtained in line 3 of Table 1. Rio de Janeiro was chosen as focus because this city was the first one to achieve the 300 cases (tenth epidemiological week). Rio de Janeiro and Duque de Caxias have high correlation with no shift (line 3, column 2); Magé has higher correlation with Duque de Caxias

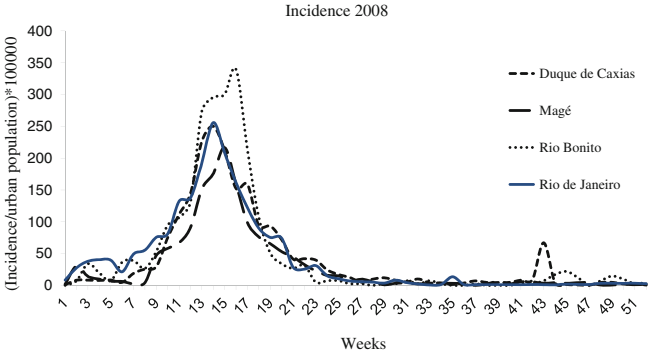


Fig. 3 Incidences of dengue cases in 2008 for cities that have correlation with Rio de Janeiro

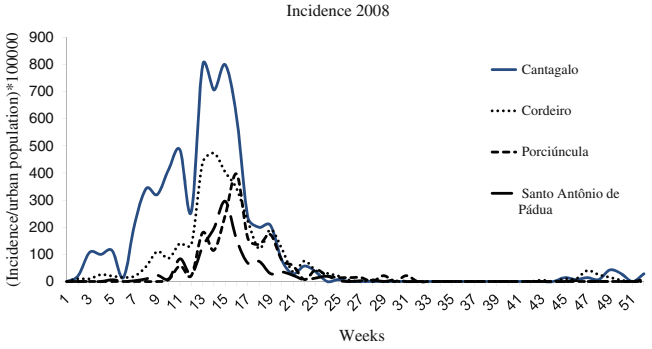


Fig. 4 Incidences of dengue cases in 2008 for cities that have correlation with Cantagalo

with a shift of 1 week and finally Rio Bonito has higher correlation with Magé with a shift of 2 weeks.

In Fig. 4 are presented the correlations obtained in line 8 of Table 1. Cantagalo was chosen as focus because this city was the first one to achieve the 300 cases (fifth epidemiological week). Cantagalo and Cordeiro have high correlation with no shift (line 8, column 2); Santo Antônio de Padua has higher correlation with Cordeiro with a shift of 1 week and finally Porciúncula has higher correlation with Santo Antônio de Padua with a shift of 2 weeks.

4 Conclusions

The hypothesis of an association between the occurrence of dengue cases between different cities in the state of Rio de Janeiro and surrounding areas was tested. The proposed methodology identified significant correlation between cities without delay,

this results suggests that the dengue epidemic occurred simultaneously in both cities, while correlations with delay may provide evidence that the mobility of people may be responsible for the spread of the disease among the regions of the state.

Using the proposed methodology, we identified the cities: Nova Iguaçu, Niterói, Rio de Janeiro, Seropédica, Campos dos Goytacazes and Cantagalo as focus of the disease in the year 2008. Then we calculate the correlations with n-delay, $n = 0, 1, 2, 3$ for the focus cities with the other cities that were selected. We were able to justify part of the significant correlations between various cities through the pendular mobility among regions of the state. The correlations that we can not explain could be independent events or characterize one diffusive process.

This information could provide an efficient control framework to guide health authorities in decision making. Once verified that dengue does not emerge at the same time in all state, and that there exist cities with potential for further spread (due to the concentration of industrial activities, market, turism, etc.) the control services could concentrate resources in a more efficient way in cities that are potential sources of spread.

Based on the identification of the propagation cascade of dengue from the focus into the other municipalities, the next step is the construction of a topological network, representing these spread dynamics coupled with human mobility data.

References

1. Adams, B.: Kapan, D.D.: Man bites mosquito: understanding the contribution of human movement to vector-borne disease dynamics. PloS one (Public library of science). **4**(8), e6763 (2009)
2. Andrade, P.G., Marques, C.: Tranformações do Território e Estruturação do Espaço no Norte Fluminense: Impactos na Dinâmica Migratória. In: XVI ENANPUR. ST1-Produção e Estruturação do espaço urbano e regional. (2015)
3. Câmara, F.P., Theophilo, R.L.G., dos Santos, G.T., Pereira, S.R.F.G., Câmara, D.C.P., de Matos, R.R.C.: Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas. Rev. Soc. Bras. Med. Trop. **40**(2), 192–196 (2007)
4. Colizza, V., Barrat, A., Barthélemy, M., Vespignani, A.: The role of the airline transportation network in the prediction and predictability of global epidemics. Proc. Natl. Acad. Sci. U.S.A. **103**(7), 2015–2020 (2006)
5. da Cruz, J.L.V., Passos, W.S.: A dimensão socioespacial do desenvolvimento na bacia petrolífera de Campos: Uma discussão do ponto de vista do mundo do trabalho. In: XIII Seminário Internacional RII. VI Taller de Editores Rier. Salvador. Brasil (2014)
6. de Farias, L.A.C.: Interações Espaciais na Rede Urbana Fluminense. Uma Análise Comparativa dos Deslocamentos Pendulares de População em: e 2010, p. 2014. Instituto de Geociências. Universidade Federal do Rio de Janeiro, Dissertação de Mestrado (2000)
7. de Medeiros Junior, H.: Desconcentração econômica e atratividade regional no estado do Rio de Janeiro entre 2000 e 2010. Cadernos do Desenvolvimento Fluminense, Rio de Janeiro. n.1 (2013)
8. de Santana, C.N., Fontes, A.S., Cidreira, M.A. dos S., Almeida, R.B., González, A.P., Andrade, R.F.S., Miranda, J.G.V.: Graph theory defining non-local dependency of rainfall in Northeast Brazil. Ecol. Complex (Elsevier). **6**(3), 272–277 (2009)
9. Gubler, D.J.: Dengue viruses: their evolution, history and emergence as a global public health problem. In: Gubler, D.J., Ooi, E.E., Vasudevan, S., Farrar, J. (eds.) Dengue and Dengue Hemorrhagic Fever, CABI (2014)

10. Kindhauser, M.K., Allen, T., Frank, V., Santhana, R.S., Dye, C.: Zika: the origin and spread of a mosquito-borne virus. Bull. World Health Org. 171082 (2016)
11. Lin, C.H., Wen, T.H.: Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. Int. J. Environ. Res. Public Health (Molecular diversity preservation, International). **8**(7), 2798–2815 (2011)
12. Liu-Helmersson, J., Quam, M., Wilder-Smith, A., Stenlund, H., Ebi, K., Massad, E., Rocklöv, J.: Climate change and Aedes vectors: 21st century projections for dengue transmission in Europe. EBioMedicine (Elsevier). **7**, 267–277 (2016)
13. Ministério da Saúde: Orientações Gerais, Prevenção e Combate, Situação Epidemiológica/Dados. In: Portal da Saúde (2016). Available via DIALOG. http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/svs/dengue. Accessed 15 June 2016
14. Mondini, A., Chiaravalloti-Neto, F.: Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. Sci. Total Environ. **393**(2), 241–248 (2008)
15. Nunes, M.R.T., Faria, N.R., de Vasconcelos, J.M., Golding, N., Kraemer, M.U.G., de Oliveira, L.F., da Silva Azevedo, R.S., da Silva, D.E.A., da Silva, E.V.P., da Silva, S.P., others.: Emergence and potential for spread of Chikungunya virus in Brazil. BMC Med. (BioMed Central). **13**(1), 1 (2015)
16. Projeção da população do Brasil e das Unidades da Federação. In: Instituto Brasileiro de Geografia e Estatísticas (IBGE) (2016). Available via DIALOG. http://www.ibge.gov.br/apps/populacao/projecao/. Accessed 15 June 2016
17. Rodriguez, I.B., Cordeiro, M.T., Braga, C., de Souza, W.V., Marques, E.T., Cummings, D.A.T.: From re-emergence to hyperendemicity: the natural history of the dengue epidemic in Brazil. PLoS Negl. Trop. Dis. Public Libr. Sci. **5**(1), e935 (2011)
18. Saba, H., Vale, V.C., Moret, M.A., Miranda, J.G.V.: Spatio-temporal correlation networks of dengue in the state of Bahia. BMC Public Health (BioMed Central Ltd). **14**(1), 1085 (2014)
19. Secretaria de Estado de Saúde do Rio de Janeiro, Subsecretaria de Vigilância em Saúde. In: Sistema de Informao de Agravos de Notificao (SINAN) (2016). Available via DIALOG. http://sistemas.saude.rj.gov.br/tabnet/deftohtm.exe?sinan/dengue.def. Accessed 15 June 2016
20. Silva, B.B.M., Miranda, J.G.V., Corso, G., Copelli, M., Vasconcelos, N., Ribeiro, S., Andrade, R.F.S.: Statistical characterization of an ensemble of functional neural networks. Eur. Phys. J. B. (Springer). **85**(10), 1–9 (2012)
21. Stoddard, S.T., Morrison, A.C., Vazquez-Prokopec, G.M., Soldan, V.P., Kochel, T.J., Kitron, U., Elder, J.P., Scott, T.W.: The role of human movement in the transmission of vector-borne pathogens. PLoS Negl Trop Dis. (Public library of science). **3**(7), e481 (2009)
22. Struchiner, C.J., Rocklöv, J., Wilder-Smith, A., Massad, E.: Increasing dengue incidence in Singapore over the past 40 years: population growth, climate and mobility. PloS one (Public library of science). **10**(8), e0136286 (2015)

The Impact of the Public-Private Investments in Infrastructure on Agricultural Exports in Latin American Countries

Bárbara Soriano and Amelia Pérez Zabaleta

Abstract Agricultural activity promotes poverty reduction. There is still an important infrastructure investment gap to enhance agricultural productivity. The public-private partnerships arise as a channel to cover it. This article analyzes the relation between public-private investment in infrastructure and agricultural exports in Latin American countries. We use a panel data sample composed by 14 countries observed over the period of 17 years, from 1995 to 2011 to which we apply panel data techniques. Results show that public-private investment in infrastructure has a positive impact on agricultural exports. The impact of private investments more than doubles the impact of public investments. The role of the public sector is crucial to guarantee the positive impacts of public-private investment on the recipient country by providing solid institutions framework and the appropriate investment climate.

Keywords Public-private partnership · Agricultural exports · Infrastructure Panel data

1 Introduction

Empirical evidence supports the view that agricultural development promotes poverty reduction [11, 43]. The positive effect of increasing productivity on poverty reduction materializes if agricultural productivity is enhanced through the integration of developing countries in global value chains [27]. Increasing productivity in agricultural sector depends on infrastructure, well-functioning domestic markets, appropriate institutions, and access to appropriate technology [31]. Physical infrastructure was found to have the greatest impact on exports [32]. Adequate infrastructures in transportation, energy and telecommunication enhance the domestic and international

B. Soriano (✉) · A. P. Zabaleta
Cátedra AQUAE de Economía del Agua (Fundación Aquae -UNED), Madrid, Spain
e-mail: barbarasoriano@gmail.com

A. P. Zabaleta
e-mail: catedraeconomiadelagua@cee.uned.es

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_22

competitiveness in developing countries [10, 50, 52]. Investment in infrastructure reduces the transaction costs in developing countries [55], it strengthens the links between local producers and consumers and it facilitates access of farmers to local and regional markets [46].

Developing countries still face large funding gaps to invest in infrastructure [44]. Infrastructure investment must rise between 1.8 and 2.3 trillion dollars per year by 2020 to meet the needs of developing countries, according to available estimates by [44]. While traditional transnational corporations remain the largest investors in infrastructure [48] new options have been settled to cover the investment needs in developing countries. The last goal of the Sustainable Development Goal (SDGs) provides for the revitalization of the global partnership for sustainable development, arguing that complex challenges require global and integrated efforts from all the stakeholders. This assessment opens new alliance opportunities for public and private sectors. The private sector plays a fundamental role facilitating investment and knowledge. The participation of the public sector is pivotal to attract the participation of the private sector, by creating adequate investment climates [52] and promoting public-private partnerships [45, 47, 48].

While the relationship between Foreign Direct Investment (FDI) and trade has been the subject of several papers, the relationship between public-private partnerships and trade has yet to be analyzed. The aim of this article is to analyze the relationship between public-private partnerships in infrastructure and the agricultural trade in developing economies. It seeks to test the hypothesis that public-private investment in infrastructure is positively related to the volume of agricultural exports. We focus the analysis on Latin American countries. Since the beginning of the century, the public-private partnership in infrastructure investments have greatly increased in Latin American countries, from 100 public-private projects in infrastructure in 2000 to more than 200 projects in 2014 [54].

The remainder of the paper is structured as follows. Section 2 reviews the literature on the relationship between investment and trade. In Sect. 3 we present the sample of countries and descriptive analysis of public-private investments in infrastructures and agricultural exports. In Sect. 4, we describe the empirical framework. The main results and discussion are summarized in Sect. 5 and in Sect. 6 we expose the principal conclusions.

2 Literature Review

The relation between investment and trade has been analyzed by the research community from different points of view. One of the key issues assessed on the relation between investment and trade is the direction of the causality relationship. There is a greater consensus about the fact that private investment precedes to trade [3, 18, 26, 30, 33]. Otherwise, [8] conclude that trade leads to higher private investment. Reference [1] suggest that there is a bidirectional relationship between trade and investment, with no clear causality in either direction.

Some authors have analyzed the positive or negative sign of the relationship between FDI and trade. There exists a complementary relationship when the private investment contributes positively to boost the exports of the recipient country. On contrary, a substitution relationship takes place when private investment has a negative impact on the exports of recipient country [14, 28].

Other authors have considered a sector based approach. Reference [41] analyze the private investment broken down according to the type of product, industry and manufacture production. They conclude that the relationship between trade and private investment varies depending on the level of disaggregation. If the analysis focuses on product and industry, investment and trade are substitutes. But, they are complements when the analysis is based on a higher disaggregation level. Reference [17, 34] study the relationship between private investment in agriculture and food trade in Canada and Sub-Saharan countries, respectively. Both studies conclude that private investment in the agricultural sector and food trade are complements.

Regarding investments in infrastructure, several studies have analyzed the role of the infrastructure on trade. Reference [32] found that the impact of physical infrastructure on exports is higher than that of other indicators as border and transport efficiency or business and regulatory environment. Reference [21] found that poor roads and ports, poorly performing customs agencies and procedures, weakness in regulatory capacity, and limited access to finance and business services affected trade. Trade facilitation by investing in physical infrastructure and regulatory reforms, improve the export performance of developing countries. Indeed, this positive impact is far important than variations in tariffs in explaining North-South trade [16]. According to the World Bank (2013), lack of proper infrastructure pushes logistics costs to as much as 25% of the food product value for Latin American countries, compared with around 9% for OECD countries.

Existing empirical studies use different data and estimation techniques to study de relationship between investment and trade. To analyze the direction and sign of the potential causal relationship between investment and trade most of the authors apply a Granger Causality test. They analyze if current and past performance of investment explains current exports or the relationship follows the opposite direction [3, 33]. Other studies use gravity models to explain bilateral trade analyzing the variables that measure the weight of the countries involved in trade (population, Gross Domestic Product –GDP– and FDI) and variables that measure the distance between them (trade barriers and language) [8, 28]. Finally, some authors broaden the sample size and analyze the relationship between investment and trade for a set of countries using panel data. Gyfalson, 1997 proposed a theoretical model where the determinants of exports are the population, GDP per capita, productive sector, inflation, dependences on primary exports, investment and economic. Reference [17] proposed that trade depends on the investment level, the exchange rate and the GDP. In the model proposed by [18] the determinants of the trade openness of a country are the GDP per capita, the inflation rate, institutional quality, macroeconomic volatility and financial openness. Reference [12] study how GDP and investment explain exports.

Finally there are several studies that analyze the factors facilitating a positive impact of the foreign investment on developing countries: (1) It is required a

minimum GDP per capita level and human capital [9, 24]; (2) a political and legal framework has been built and supported by multilateral agencies, labor organizations, non-governmental organizations and civil society groups [29]; and (3) There must exist a developed financial markets to transmit the impact of investments to national incomes [2]. Reference [25] find that a legal and regulatory framework that protects the rights and the obligations of the investors is decisive to target foreign investment to developing countries. As institutions framework is more robust, the foreign investments are greater [7, 37]. Reference [53] concludes that strengthening regulations and institutions will be even more important to boost agricultural trade, especially when compared to other sectors.

3 A First Look at the Data

The aim of the study is the analysis of the relationship between public-private partnerships in infrastructure and agricultural exports. With this purpose, we use data on agricultural trade from World Development Indicators Bank Database. The database provides information about the agricultural raw material exports expressed as percent of merchandise exports and the merchandise exports expressed in current US$. We obtain the agricultural raw material exports expressed in US$ by multiplying both variables.

Concerning investments projects in infrastructure, we use the database of the Public-Private Infrastructure Advisory Facility (PPIAF). This initiative gives support to developing countries to create adequate investment environments (policy guidance, development of regulation, consolidation of institutions and governance) that encourage foreign investors to invest in sectors not covered by the public sector. The projects cover investment in transport (roads, bridges, tunnels, terminals and dredging of channels projects), telecommunications (investment in fixed access network and mobile communications), energy (generation, transmission and distribution of electricity and natural gas) and water and sanitation (water transport systems, water treatment and sewerage plants and water and sanitation services). The database provides information about the public-private investments in infrastructure (current US $). Furthermore, it splits the investment information into two parts: (1) private investment and (2) public investment.

The sample of the study is made up of 14 Latin American countries (Table 1), over the period 1995–2011.

Regarding to the public-private investments in infrastructure Fig. 1 shows that it stood at 50 billion dollars at current prices in period 1995–1999 in developing countries. This figure has increased up to 140 billion dollars in real terms in 2010–2012 [54]. There is a clear targeting of the public-private investment into the energy sector, reaching more than half of total public-private investment during the whole period.

Table 1 Sample of Latin American countries

Countries of the sample	Region
Bolivia	Andean
Colombia	Andean
Ecuador	Andean
Peru	Andean
Venezuela	Andean
Brazil	Brazil
Argentina	South Cone
Chile	South Cone
Paraguay	South Cone
Costa Rica	Mesoamerica
Guatemala	Mesoamerica
Honduras	Mesoamerica
Nicaragua	Mesoamerica
Mexico	Mexico

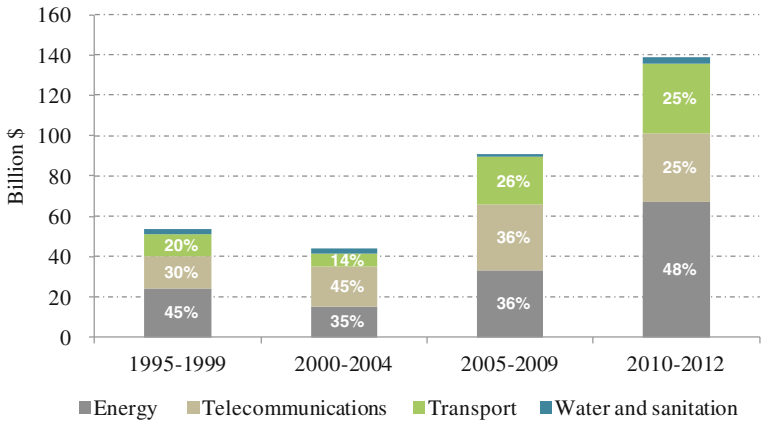


Fig. 1 Evolution of the public-private investment in infrastructure by investment sector. *Source* [39]

Energy is a key element in the development process. It is required for food processing, transportation, fertilizer production and use of industrial equipment, among other multiple uses [40]. Investments in communications also contribute to development by providing greater access to timely information (prices, clients, suppliers), by enhancing the bargaining power of small farmers, and increasing trade and agriculture production [22, 23]. The availability of adequate transportation infrastructure facilitates access of farmers to markets [51]. It is worth to mention, that public-private investment in water and sanitation represents less than 5% of the public-private

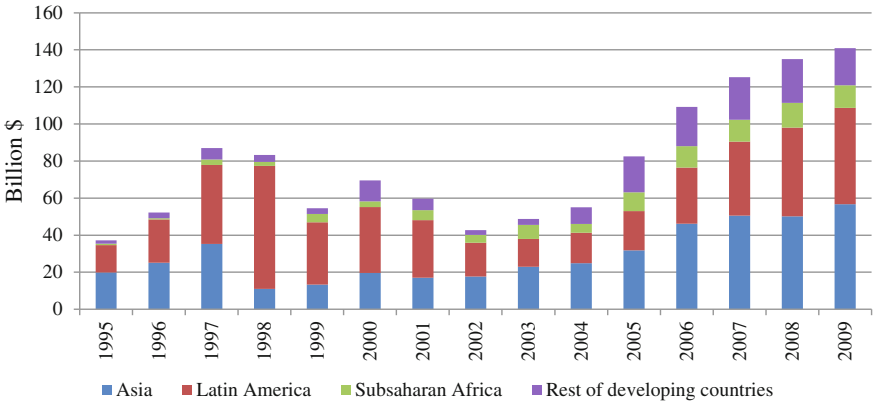


Fig. 2 Public-Private investments in infrastructure by region, period 1995–2009. *Source* Own elaboration based on [38]

investment in infrastructure over the period of study. Concerning the structure of the public-private partnerships, foreign private participation clearly leads the investment in infrastructures. It reaches more than 90% of total investment in 2012 [54]. Latin America is the region that accumulates greater investments in infrastructures, reaching 52 billion dollars in 2009 (Fig. 2).

Regarding trade, the dollar value of world merchandising trade has increased by more than 7% per year on average over the last twenty years (1980–2011). Food trade shows similar trend. In last forty years, the number of calories exchanged through the global food trade has multiplied fivefold [13].

Latin American countries have contributed highly to global agricultural production and trade. While there are significant differences between countries, the region is overall a net agricultural exporter. Exports of agricultural products have grown at about 8% annually since the mid-90s. It represents 13% of agricultural trade, up from 8% in the mid-90s [53].

Figure 3 shows a first sight of the potential relationship between agricultural exports and public-private investments in Latin American Countries. The agricultural exports increased in all the Latin American regions, reaching between 2 and 5 billion dollars per year in the period 2006–2011. Brazil and the South Cone are the largest exporters of agricultural products in Latin America. Public-private partnerships decreased between during the first five years of the period, except Mexico. Since 2000, the public-private investments also increased in all the regions. It is worth to highlight the public-private investment in infrastructure in Brazil. It is nearly ten times the public-private investments in Andean Region or South Cone in the period 2006–2011.

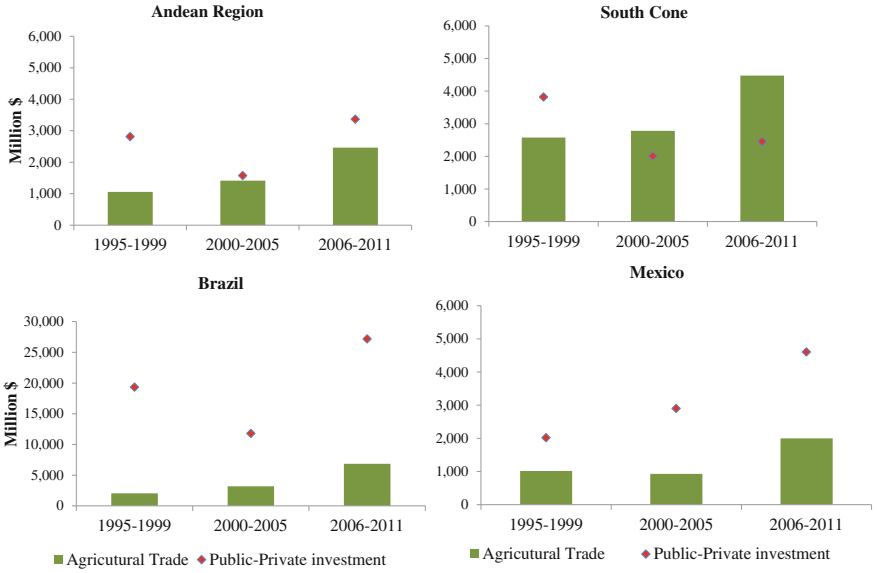


Fig. 3 Agricultural exports and public-private investments in infrastructure by areas of the Latin American region, period 1995–2011. *Source* Own elaboration

4 Econometric Set-Up and Empirical Results

4.1 Methods

We build our model on the base of previous causality studies that suggest that the investment precedes trade [3, 26, 30, 33]. Reference [18] explained why investment precedes trade by arguing that investment is often directed towards the sector of traded goods. Investments in infrastructure also precede trade because it may increase traded food output. Reference [42] concludes that one of the fundamental economic factors affecting international trade is investment (considering technology and energy). Investment in physical infrastructure can facilitate the integration of fresh players into international supply chains. It can also contribute to a reduction of transportation costs.

In our model, we explain agricultural exports using public and private investments in infrastructure as explanatory variables. We consider the public-private investment splits into two parts: (1) the private investment and (2) the public investment. So, we consider that this split of investment constitutes a contribution of our paper to the literature. So, we try to test the hypothesis that there is a positive relationship between public and private investments in infrastructure and agricultural exports. As control variables we consider in the model GDP per capita of the exporter country [12, 18]

and the nominal exchange rate [17]. All are annual variables and are expressed in logarithms.

The estimation of the baseline model (1) is as follows:

$$\text{L\_Agri\_X}_{it} = \alpha_0 + \alpha_1 \text{L\_Private\_invest}_{it} + \alpha_2 \text{L\_Public\_invest}_{it} + \alpha_3 \text{L\_GDP\_cap}_{it} + \alpha_4 \text{L\_XRT}_{it} + \varepsilon_{it}$$

where,
$i = 1 \cdots 14$ Latin American countries;
$t = 1 \cdots 17$ years (period 1995–2011);
L_Agri_X_{it} = Logarithm of agricultural exports, country i, year t;
$\text{L\_Private\_Invest}_{it}$ = Logarithm of private investment in infrastructure, country i, year t;
L_GDP_cap_{it} = Logarithm of GDP per capita, country i, year t;
$\text{L\_Public\_Invest}_{it}$ = Logarithm of public investment in infrastructure, country i, year t;
L_XRT_{it} = Logarithm nominal annual exchange rate, country i, year t;
ε_{it} is the error term;

We apply panel data techniques to a sample with two dimensions (countries and time) The main advantage of the panel data is the ability to control for unobserved heterogeneity in cross-sectional models [5]. Applying panel data requires selecting one of the following two specifications: (1) a model allowing each cross-sectional unit (or group of cross-sectional units) to have its own intercept. It assumes that each cross-sectional unit has a non-stochastic group-specific component. The unobservable effects are controlled by including N dummy variables, one for each country. It is known as the Fixed Effects model (FE). One potentially significant limitation of fixed effects models is that they cannot be used to investigate the effects of time-invariant dependent variables. (2) a model considering that each cross-sectional unit has a stochastic unobserved effect. The unobservable effects are treated as a component of the random error term. It is named Random Effects model (RE). There is a common problem of bias in the estimates due to correlated effects [4].

We run the Hausman test where the null hypothesis is that the preferred model is RE versus the alternative the FE [19, 20]. The RE estimates are consistent and efficient under the null while the FE estimates are not efficient. Another way to look at the results of the test is whether the errors are correlated with the regressors. The null hypothesis is that they are not, i.e., the test does not reject the RE specification. The Hausman test shows that we should base our inference on the results of the FE specification (Table 2).

We do so by introducing in the model a classification of the countries according to the quality of the institutional framework (model 2), instead of a dummy for each country. Improvements in institutional quality are a source of comparative advantage and may promote trade [42]. Institutions refer to the social norms, ordinary laws, political regimes or international treaties, within which policies are determined and

Table 2 Categorization of countries according to the institutional quality and investment climate

Average value of political regime over the period	Stability indicator	Description	Countries
Lower/Equal to 6	1	**Low** institutional quality and investment climate	Peru and Venezuela
Higher than 6 and lower/equal to 8	2	**Medium** institutional quality and investment climate	Argentina, Bolivia, Brazil, Colombia, Ecuador, Guatemala, Honduras, México, Nicaragua and Paraguay
Higher than 8 and lower/equal to 10	3	**High** institutional quality and investment climate	Chile and Costa Rica

Source Own elaboration

economic exchanges are structured. The political regimen has been proposed previously as a proxy variable by other authors [1, 18]. This variable ranges from −10 (autocracy) to 10 (full democracy). We assume that democratic countries provide higher quality institutions and stable investment climates than autocratic countries. Based on its values, we define a time-invariant indicator that ranges from 1 to 3, by classifying countries according to the average value of political regime over the period as follows:

We test the assumption of the stochastic disturbance term in model (1) and model (2). We apply the Wooldridge test to identify the existence of serial correlation in the errors [49]. The test result shows that there is serial correlation. We apply Wald test revealing the existence of heteroskedasticity problems [15]. To correct for correlation and heteroskedasticity, we apply the panel corrected standard errors (PCSE) to models (1) and (2). PCSE assumes that the disturbances are by default heteroskedastic and contemporaneously correlated across units [6].

The definition of the variables and descriptive statistics for the variables included in the specifications are summarized in Table 3.

5 Results and Discussion

The results of specification (1) are summarized in Table 4. As it can be seen, the coefficient of the private investment in infrastructure is positive and statistically significant. This result suggests that this factor enhances agricultural exports in developing countries. Our result agrees with previous results that find a positive relationship between FDI and trade [3, 14]. The coefficient indicates that 1% increase in private investment in infrastructure would generate an increment of 0.23% of the agricultural exports.

Table 3 Measures and descriptive statistics

Variable	Measure	Source	Obs	Mean	Std. dev.	Min	Max
Agri_ export	Current million US$	World Development Indicators	234	773	1,310	9	9,044
GDP per capita	Current US$ per habitant	World Development Indicators	238	3,943	2,865	699	14,501
XRT	Local currency per US$	World Development Indicators	233	1,394	4,428	0	25,000
Private_ invest	Current million US$	PPI World Bank	210	1,966	5,160	2	33,292
Public_ invest	Current million US$	PPI World Bank	105	800	2,066	0	15,107
Stability		Polity IV Project Database	238	2	1	1	3

Source Own elaboration

Reference [12] found that an increase in 1% of foreign direct investment will lead to an increase of 0.64% on exports.

If we look at the coefficient of the public investment in infrastructure we can see that it is positive and statistically significant. It means that an increment in public investment in infrastructure generates an increment of the agricultural exports. This result is in accordance with the results obtained by [35, 36]. They found a positive impact of public investments in transport infrastructures. They conclude that providing key infrastructures, public sector can create synergies with private sector. Private investments that were previously uneconomic become profitable and the country may become a potential market for investment.

Comparing the coefficients of the private investments and public investment in infrastructures we see that the coefficient of the private investment in infrastructure triples the coefficient of the public investment in infrastructure. It implies that the impact of private investment of agricultural exports is higher than the impact of public investment.

The results of specification (2) show that the coefficients of the levels of the quality of institutions and investment climate variable are positive and statistically significant. It means that the investment climate promoted by governments channels the impact of investment in infrastructure on agricultural exports. The better is the investment climate, the higher is the impact of investments in infrastructures on

Table 4 Results of the regression model

Dependent variable: agricultural exports		
Variable	Model 1	Model 2
	Base model (PCSE)	Base model stability fixed effect (PCSE)
Private_ Invest	0.232***	0.169***
	(0.055)	(0.047)
Public_ Invest	0.062*	0.068*
	(0.029)	(0.027)
GDP per capita	1.334***	1.332***
	(0.139)	(0.147)
XRT	0.136***	0.093***
	(0.028)	(0.025)
Stability investment climate		
Medium		0.733*
		(0.348)
High		0.977*
		(0.397)
Intercept	−0.077	−0.22
	(0.93)	(1.13)
N	102	102
R^2	0.93	0.95
Hausman statistic	31.90	32.69
	(0.00)	(0.00)

Note: *, **, ***, denotes statistical significante level at 5, 1 and 0.1%
Figures in parentheses are the coefficients standard errors.
Source Own elaboration.

agricultural exports. This result is consistent with previous studies that concluded that the investment climate is key to attract private investment [1, 7, 18, 37, 42, 53].

The coefficient of the private investment remains higher than the coefficient of the public investment, although the gap is lower than in the in base specification. These results show that government contributes to boost agricultural exports not only by investing in infrastructures but also by making the investment scenarios more attractive.

GDP per capita has the highest statically significant coefficient. As GPD per capita grows, the export capacity of developing countries improves. Reference [12] found that economic growth is elastic to exports. They found that an increase in 1% on GDP will lead to an increase in 1.37% on exports. Finally, the exchange rate coefficient shows that the devaluation of the national currency contributes to increase agricultural exports positively. These results are consistent with those of [1, 17].

6 Conclusions

There is still an important infrastructure investment gap to boost agricultural activity as the main source of livelihood for the poor population. The public-private partnerships arise as a channel to cover this gap.

The goal of the article was to analyze the relationship between public-private investments in infrastructure and agricultural exports in developing countries. We test the hypothesis that the public-private investment in infrastructure has a positive impact on the agricultural exports, using a panel data sample of 14 Latin American countries covering 17 years (1995–2011).

The results provide evidence, first, about the positive contribution of public-private partnerships to boosting the agricultural exports in developing countries. Second, the impact of the private investment in infrastructure more than doubles the impact of the public investment. These results reinforce the relevance of the private sector in reducing the investment gap. The role of the public sector in the public-private partnership is also pivotal. Governments provide the investment climate that guarantees the positive impacts of the investment on the recipient country. As the investment climate is more stable, the impact of the public-private partnership is higher.

References

1. Aizenman, J., Noy, I.: FDI and trade-two way linkages? Q. Rev. Econ. Financ. **46**(3), 317–337 (2005)
2. Alfaro, L., et al.: FDI and economic growth: the role of local financial markets. J. Intern. Econ. **64**(1), 89–112 (2004)
3. Alguacil, M.T., Cuadros, A., Orts, V.: Foreign direct investment, exports and domestic performance in Mexico: a causality analysis. Econ. Lett. **77**, 371–376 (2002)
4. Allison, P.: Fixed Effects Regression Models. SAGE Pulications Inc, Thousand Oaks (2009)
5. Arellano, M.: On the testing of correlated effects with panel data. J. Econom. **59**, 87–97 (1993)
6. Beck, N., Katz, J.N.: What to do (and not to do) with time-series cross-section data. Am. Econ. Rev. **89**(3), 634–647 (1995)
7. Bénassy-Quérê, A., Coupet, M., Mayer, T.: Institutional determinants of foreign direct investment. World Econ. **30**(5), 764–782 (2007)
8. Bezuidenhout, H., Naudé, W.: Foreign Direct Investment and Trade in Southern African Development Community. UNU-WIDER Research paper, (2008/88) (2008)
9. Blomsrtöm, M., Lipsey, R.E., Zejan, M.: What explains developing country growth? NBER Working Paper Series. National Bureau of Economic Research Cambridge (1992)
10. Cervantes-Godoy, D., Dewbre, J.: Economic Importance of Agriculture for Poverty Reduction. OECD Food, Agriculture and Fisheries.Working Paper, (23) (2010)
11. Dorward, A., et al.: A Policy Agenda for Pro-Poor Agricultural Growth. ADU Working Paper 02/02 (2001)
12. Dritsaki, M., Dritsaki, C., Adamopoulos, A.: A casual relationship between trade, foreign direct investment andE and economic growth for Greece. Am. J. Appl. Sci. **1**(3), 230–235 (2004)
13. FAO: Safeguarding Food Security in Volatile Global Markets. FAO, Rome (2011)
14. Fontagné, L.: Foreign Direct Investment and International Trade. Complements or Substitutes? OCDE Science, Technology and Industry Working Papers, 03, p. OECD Publishing (1999)

15. Fox, J.: Applied Regression Analysis, Linear Models, and Related Methods. Sage Publications Inc, Thousand Oaks (1997)
16. Francois, J., Kepler, J., Manchin, M.: Institutions, Infrastructure, and Trade. Policy Research Working Paper. World Bank. WPS4152 (2007)
17. Furtan, W., Holzman, J.: The Effect of FDI on Agriculture and Food Trade: An Empirical Analysis. Agricultural and Rural Working Paper Series, 21-601-MIE(068) (2004)
18. Ghosh, I.: The relation between trade and fdi in developing countries – a panel data approach. Glob. Econ. J. **7**(3), 1–32 (2007)
19. Greene, W.H.: Econometric Analysis, 6th edn. Pearson Prentice Hall, New York (2007)
20. Hausman, J.A.: Specification test in econometrics. Econom. J. Econom. Soc. **46**(6), 1251–1271 (1978)
21. Hoekman, B., Nicita, A.: Trade policy, trade costs, and developing country trade. World Dev. **39**(12), 2069–2079 (2011)
22. IICD: Las TIC para el sector agrícola. Impacto y lecciones aprendidas de programas apoyados por el IICD (2006)
23. IICD: Accelerating development. Building capacities through innovation. Annual Report. IICD, People ICT-Development (2012)
24. Jyun-yi, W., Chih-Chiang, H.: Does foreign direct investment promote economic growth?? evidence from a threshold regression analysis. Econ. Bull. **15**(12), 1–10 (2008)
25. Lamech, R., Saeed, K.: What International Investors Look For When Investing In Developing Countries. Results form a survey of international investors in the power sector. Energy and mining sector board discussion paper, no. 6. The World Bank Group. The Energy and Mining Sector Board (2003)
26. Liu, X., Wang, C., Wei, Y.: Causal links between foreign direct investment and trade in China. China Econ. Rev. **12**, 190–202 (2001)
27. Maertens, M., Colen, L., Swinnen, J.F.M.: Globalisation and poverty in Senegal: a worst case scenario? **38**, 31–54 (2011)
28. Magalhaes, M., Africano, A.P.: A panel analysis of FDI impact on International Trade. NIPE. Working papers series, WP6/2007 (2007)
29. Moran, T.H.: Foreign Direct Investment and Development. Peterson Institute For International Economics (2011)
30. Pacheco-López, P.: Foreign direct investment, exports and imports in Mexico. World Econ. **28**(8), 1–24 (2005)
31. Pinstrup-Andersen, P., Shimokawa, S.: Rural Infrastructure and Agricultural Development (2006)
32. Portugal-Perez, A., Wilson, J.S.: Export performance and trade facilitation reform: hard and soft infrastructure. World Dev. **40**(7), 1295–1307 (2012)
33. Pramadhani, M., Bissoondeeal, R., Driffield, N.: FDI, Trade and Growth, A Casual Link? Aston University, Birmingham (UK) (2007)
34. Rakotoarisoa, M.A.: A Contribution to the Analyses of the Effects of Foreign Agricultural Investment on the Food Sector and Trade in Sub-Saharan Africa. FAO commodity and trade Policy Resesarch Working Paper, 33 (2011)
35. Ronald-Holst, B.: Infrastructure as a catalyst for regional integration: Scenario analysis for Asia. Conference paper for the Asian Development Bank (2006)
36. Scandizo, S., Sanguiretti, P.: Infrastructure in Latin America: Achieving high impact management. Discussion draft on the 2009 Latin America emerging market forum (2009)
37. Shleifer, A., Wolfenzon, D.: Investor protection and equity markets. J. Financ. Econ. **66**, 3–27 (2002)
38. Soriano, B., Garrido, A., Novo, P.: Agua virtual y cooperación internacional. Las relaciones entre el comercio de agua virtual y la Ayuda Oficial al Desarrollo en la Cooperación Internacional. Madrid, Spain: Fundación Canal, 205 (2013)
39. Soriano, B., Garrido, A.: The role of private sector in development: The relation between public-private investment in infrastructure and agricultural exports in developing countries. Economía Agraria y Recursos Naturales **15**(2) (2015)

40. Stout, A.: Handbook of Energy for World Agriculture. Elsevier, Amsterdam (1990)
41. Swenson, D.L.: Foreign investment and the mediation of trade flows. Rev. Int. Econ. **12**, 609–629 (2004)
42. Tang, H.: World Trade Report 2013 - factors shaping the future of world trade. World Trade Rev. **13**(04), 733–735 (2014)
43. Thirtle, C., Irz, X.: The relationship between changes in agricultural productivity and the incidence of poverty in developing countries. Paper prepared for DFID, Imperial Collegue, London (2014)
44. UN: The role of public investment in social and economic development. Public Investment: Vital for Growth and Renewal, but should it be a Countercyclical Instrument. United Nations, New York and Geneva (2009)
45. UN: Report on the Ministerial Meeting: Enhancing the mobilization of financial resources for least developed countries' development. United Nations Office of the High Representative for the Least Developed Countries, Landlocked Developed Countries and Small Island Developing Countries, Lisbon (2010)
46. UN: General Assembly resolution 65/220. The right to food, United Nations (2011)
47. UNCTAD: World investment report. UNCTAD (2011)
48. UNTT: Challenges in raising private Sector resources for financing sustainable development. UNTT Working Group on Sustainable Development Financing, pp. 1–28 (2013)
49. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge (2002)
50. World Bank: World Development Report. International Development Trends. Agriculture and Economic Development, World Development Indicatos (1982)
51. World Bank: Agricultural Investment Sourcebook. Agriculture and Rural Development, The International Bank for Reconstruction and Development. The World Bank (2005)
52. World Bank: World Development Report. Agriculture for Development. Overview. The World Bank, Washington, DC (2008)
53. World Bank: Future looks bright for food production in Latin America and Caribbean (2013)
54. World Bank: Private Participation in Infrastructure Project Database. http://ppi.worldbank.org (2014). Retrieved January 2014
55. WTO: World Trade Report. Factors shaping the future of world trade, World Trade Organization (2013)

Major Simulation Tools for Biochemical Networks

Gökçe Tuncer and Vilda Purutçuoğlu

Abstract As biochemical networks become more popular, the number of simulation tools grows rapidly. Although most of the tools have similar functionality, they differ in their algorithms and capabilities. Here, we present the major simulation tools applied in biochemical networks and describe their supported algorithms with details. We consider that the capacities of each tool in terms of the simulation, inference or visualization of the different types of biological networks, their supported algorithms and the features of these algorithms as well as the mathematical background in all these calculations can be helpful for the researchers when they choose the most appropriate tool for their analyses.

Keywords Simulation tools · Biochemical systems · Deterministic simulation methods · Stochastic simulation methods

1 Introduction

As the development of the technology for producing data, the researcher can get more information about the biological processes which constitute a network or system in the end. All biological activations are directed by different networks which are denoted by nodes, i.e., genes or species of interest, and the edges, i.e., interactions between the species.

In laboratories, even though the capacities of the production of data increase day by day, the researchers can prefer the simulation of the systems in advance of their experiments due to the cost of these experiments and the complexity of the actual biological processes. In fact, the simulation of the networks can detect biologically

G. Tuncer · V. Purutçuoğlu (✉)
Department of Statistics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: vpurutcu@metu.edu.tr

G. Tuncer
e-mail: tuncer.gokce@metu.edu.tr

A. A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics III*, Springer Proceedings in Mathematics & Statistics 224,
https://doi.org/10.1007/978-3-319-74086-7_23

more interesting questions as it helps us to observe the plausible behaviour of the system beforehand.

Hence, there are a number of tools developed for this purpose. In this study, we choose Cellware, COPASI, Dizzy, Dynetica, E-CELL, GENESIS, Jarnac combined with JDesigner, System Biology Toolbox and Virtual Cell due to their common applications among their alternatives. All these tools are mainly designed for the deterministic simulation. Whereas most of them can also support the stochastic simulation algorithms and some of them can infer the model parameters too. In the organization of the chapter, we initially explain each of these tools under certain criteria and then present brief descriptions of all the supported deterministic and stochastic methods within these tools. Finally, we summarize the findings in Conclusion part.

2 Simulation Tools

In the literature of system biology, there are many tools developed by distinct groups or researchers in order to simulate the biological networks under various assumptions. Some of these tools are designed and used by particular groups. Here, we choose the most common ones among many alternatives. The selected tools are Cellware, COPASI, Dizzy, Dynetica, E-CELL, GENESIS, Jarnac combined with JDesigner, System Biology Toolbox and Virtual Cell. Below we explain them in details according to their selected versions which are denoted by the parentheses in the associated subtitles.

2.1 Cellware (3.0.1)

Cellware [1] is a multi-algorithmic environment for modeling and simulating biochemical reactions in a cell. It presents an integrated environment for various mathematical representations with a user-friendly graphical interface and a high capacity in the application of the complex systems. This tool is also known as the first grid-based modeling and simulation tool for cellular processes in systems biology.

On the other hand, this tool is written in Java and uses a proprietary file format storing information compatible to the model and the simulation. Moreover, it supports the import and export of models from the System Biology Mark-up Language (SBML) which is a free representative XML based format for interchanging different biological processes [2].

Additionally, Cellware is an open source software for only academic users. Furthermore, it has a diagrammatic interface enabling the users to draw diagrams in place of writing chemical equations manually. Accordingly, the rate expressions and parameters for the systems of interest can be entered by dialog boxes.

The main functionalities of this tool can be listed as follows [3] and the mathematical details about its supported algorithms are given in the following parts.

- **Biologist - Friendly User Interface (UI)**: User interface of Cellware is designed in the drag-and-drop format which enables the researchers to manipulate the system elements within the same or different simulations.
- **Grid Enabled**: The grid environment of the tools helps the users to implement it for large biological systems.
- **Multi-platform**: Cellware is a multi-platform software developing in Java. Hereby, the grid version implements the front-end in Java and the back-end in C++. It is also supported by Windows, Linux/Unix and Mac OS X platforms.
- **Extensive Algorithms Library for Simulation**: Cellware has an extensive library having both deterministic methods (Euler Forward and Backward method, Trapezoidal method, Explicit 4th order Runge–Kutta method, Rosenbrock method, Advanced ODE Solver(Adams–Bashforth)) and stochastic methods (Gillespie's Direct method, Gibson Next Reaction method, Explicit Time-step (also called τ-leap) method). The tool also offers the hybrid stochastic approach (StochODE method) in simulation.
- **Inference of Model Parameters**: Cellware can perform the estimation of the model parameters based on deterministic approaches by implementing the Particle SWARM algorithm which is one of the well-known optimization approaches.
- **Network Analysis**: The user can easily extract topological information from the model such as stoichiometry matrix, network statistics, independent cycles, paths, transversal nodes and conserved pathways.

Lastly, Cellware has a detailed manual and tutorial which can be obtained from its website that is regularly updated [4].

2.2 COPASI - A Complex Pathway Simulator (4.14 -Build 89)

COPASI [5] is one of the popular user-friendly softwares for the simulation, analysis and the parameter estimation of the biochemical networks. In general, this tool offers diverse analyses' methods to the researchers including the steady state, i.e., deterministic, stoichiometric state, metabolic control and the sensitivity analyses.

COPASI's native file format is XML and the documentation of its schema is available for other tools so that they can be read or written easily. COPASI also reads GEPASI (First Version of COPASI) [6] files. Moreover, COPASI is SBML [2] compatible in such a way that this software can read and write the SBML files through the `libsbml` package.

Finally, COPASI is an open source software which is free for both academic and commercial users. It has a dialog interface which eliminates the requirement for writing kinetic equations explicitly. Hereby, the user only specifies the chemical equations and rate expressions for each reaction and then states the compartments. In the end, it can estimate the parameter of the system.

On summary, the functionalities of COPASI can be listed as below [7]. The mathematical details about the associated algorithms are presented in the following parts.

- **User Interface (UI)**: COPASI has a user-friendly graphical interface created with a cross-platform application framework, called Qt, which is widely used for developing application software and allows implementing the software on all platforms supported by Qt.
- **Different Executable Versions**: COPASI has two different executable versions, namely, a graphical user interface, shortly GUI, via CopasiUI and a command line version without a GUI for batch via CopasiSE. For instance, if the users do not interact with the software, they can choose CopasiSE for their calculations.
- **Multi-platform**: The software is written in C++ to have a fast reliable simulation and software supported by the platforms such as Windows, Linux, Mac OS X and Sun Microsystems Solaris.
- **Different Simulation Methods**: COPASI uses Livermore Solver for the ordinary differential equations (ODE) method which is a part of the `ODEPACK` library for deterministic simulations and uses the Householder QR factorization for finding conservation relations in order to reduce the dimension of the systems. For stochastic simulations, the Next-Reaction method [8] is available. Furthermore, COPASI can offer hybrid simulation methods which are the deterministic numerical integrations of ODEs with the stochastic simulation algorithms. For this manner, it applies the Runge–Kutta methods with different orders. Finally, to compute the steady-state position of the systems in the ODE-based models, COPASI uses the `LAPACK` library.
- **Diverse Inference Methods**: There are several methods to estimate and simulate the model parameters deterministically. These approaches define an objective function based on the parameters and put constraints regarding the natural features of these terms. Finally, the underlying objective function is either minimized or maximized with respect to the rule of the selected optimization methods. In the optimization, COPASI can implement different techniques such as the Evolutionary Programming, Evolutionary Strategies by using Stochastic Ranking, Genetic Algorithm, Genetic Algorithm by using Stochastic Ranking, Hooke and Jeeves, Levenberg–Marquardt, Nelder–Mead, Particle Swarm, Praxis, Random Search, Simulated Annealing, Steepest Descent and the Truncated Newton methods.
- **Diverse Computation Tasks**: Besides simulation and optimization, the tool also offers methods for the steady-state calculation, metabolic control analysis, Lyapunov exponents calculation, sensitivity analysis and the time scale separation computation.

Lastly, COPASI has a detailed manual and tutorial wizard which can be found on its website regularly updated [9].

2.3 *Dizzy (1.11.4)*

Dizzy [10] presents a software tool for modeling the homogeneous kinetics of integrated large-scale genetic, metabolic and the signaling networks. The tool offers both deterministic and stochastic simulation methods. Moreover, the features of Dizzy include a modular simulation design, reusable modeling elements, complex kinetic rate laws, multi-step reaction processes, steady-state noise estimation and the spatial compartmentalization.

Dizzy is written in Java and a set of Java packages constitutes a Java application programming interface (API) for the software. Dizzy is also SBML compatible and can import/export files by using the `SBMLReader` package. It can also display models graphically by using the Cytoscape software system.

Additionally, Dizzy is an open source software which is free for both academic and commercial purposes. Furthermore, it has a simple textual interface which requires the description of the model in the textual form.

Hereby, the functionalities of Dizzy can be listed as follows [11]:

- **User Interface (UI)**: Dizzy has an intuitive and a friendly graphical user interface.
- **Simple Syntax**: The mathematically inclined users, who may want to write the kinetic equations directly, can use Dizzy easily.
- **Multi-platform**: The software is written in Java and runs on many platforms such as Windows, Linux, and Mac OS X.
- **Different Simulation Methods**: Dizzy provides a collection of deterministic (5th Order Runge–Kutta with a Fixed Step-size or an Adaptive Step-size Controller, 5/4 Dormand-Prince ODE Solver with Adaptive Step-size Controller, Implicit-Explicit ODE Solver with Step Doubling) and stochastic methods (Gibson–Bruck method and Gillespie method as the exact stochastic simulation algorithms and Gillespie time-step method as the approximate stochastic simulation algorithm) for solving the dynamics of a model. The latest version of Dizzy also supports the solver for stochastic differential equations.
- **Modular Design**: This structure enables the user to perform the ODE based solvers for the optimization that models the systems deterministically.

Finally, Dizzy owns a detailed manual and tutorial which can be found on its website [12].

2.4 *Dynetica (1.2 Beta)*

Dynetica [13] is a user-friendly modeling interface designed for the construction, visualization and the analysis of kinetic models of biological systems as well as genetic networks. It offers both deterministic and stochastic simulation methods as well as the sensitivity analysis.

Dynetica is written in Java and supports import and export of models from the SBML file format providing exchangeability of mathematical models into the biological processes.

Furthermore, Dynetica is an open source software which is free for both academic and commercial users. It has a diagrammatic interface enabling the user to draw diagrams in place of writing chemical equations manually. Accordingly, rate expressions and parameters can be entered by dialog boxes.

Finally, the functionalities of this tool can be stated as follows [14]:

- **Genetic Networks**: By means of Dynetica, it is easy to represent a genetic network which is treated as a special reaction network containing one or more genomes.
- **Multi-platform**: Dynetica is a multi-platform software developing in Java. It is also supported by the platforms such as, Windows, Linux, and Mac OS X.
- **Different Simulation Methods**: Dynetica provides the 4th Order Runge–Kutta method for the deterministic simulations and uses the Gillespie method for stochastic simulations of biochemical networks.

Lastly, Dynetica has only a manual which can be found on its website [15] and the mathematical details of the supported algorithms are shortly described in the following parts.

2.5 E-CELL (3)

E-CELL [16] shows a modeling and simulation environment for biochemical and genetic processes. Here, the users can define functions of proteins, their interactions with DNA, and further cellular metabolism functions.

Moreover, E-CELL is an object-oriented programming language written in C++ for simulating molecular processes and fully compatible with SBML.

Furthermore, E-CELL is an open source software which is free for both academic and commercial purposes. Finally, it has a dialog interface.

Main functionalities of E-CELL can be listed as follows [16–18]:

- **Integrative Simulation**: E-CELL represents an integrative environment for biochemical and genetic simulations by linking the gaps between metabolic pathways.
- **Multi-platform**: E-CELL is available in Windows and Linux.
- **Different Simulation Methods**: The users can select several ODE solvers (Euler method and Runge–Kutta method), stochastic methods (Gillespie method and Gibson–Bruck method), a discrete time simulator and a hybrid dynamic/static pathway simulation method in their calculations. Furthermore, by using the parallel computations, the tuning of parameters, metabolic control and the bifurcation analyses can be conducted in this tool. The mathematical details of these algorithm are represented in the following parts.

Finally, E-CELL has a manual and a tutorial which can be found on its regularly updated website [19].

2.6 GENESIS - The GEneral NEural SImulation System (2.4 Beta)

GENESIS [20] is a simulator for biochemical networks and neuronal systems. More specifically, it is designed for generating realistic models varying from subcellular components and biochemical reactions to complex models of single neurons and large networks. Furthermore, it can perform systems-level models and conduct neuronal analyses.

On the other side, the graphical user interface of GENESIS is the X-windows Output and Display Utility for Simulations (XODUS) written in C. But, it is not compatible with SBML.

Additionally, it is an open source software which is free for both academic and commercial users. Moreover, it has a textual interface which requires the description of the model in the textual form.

Finally, the major functionalities of GENESIS can be presented as below [20, 21]:

- **Building-block Approach**: The design of GENESIS and the interface are created with a building-block approach. With the help of this object-oriented approach, the users can easily exchange and reuse models and model components or can extend the functionality of the tool by adding new commands or simulation components to the simulator, without modifying the base code.
- **Powerful Script Language**: The script language of GENESIS is very powerful in the sense that only few lines of script are required even for a complicated simulation.
- **Multi-platform**: GENESIS runs under the UNIX-based systems with the X-Window System, including Linux, OS/X and Windows with Cygwin.
- **Inference of Model Parameters**: This tool uses the Parallel Genetic Algorithms [22] and the Parallel Simulated Annealing [23] method in the optimization when the systems are deterministically generated. It also uses the stochastic parameter search method for the stochastically generated systems.
- **Using with Kinetikit (11)**: Kinetikit [21], which is an interface and a simulator for biological signaling pathways, can be used to generate the behaviour of a system. In Kinetikit, the Exponential Euler method or the Runge–Kutta method is offered for deterministic modeling. Furthermore, it can implement the Mixed Stochastic method, Gibson–Bruck method, Gillespie method and the First Reaction method for the stochastic simulations.

In the end, GENESIS presents a detailed manual and tutorial which can be found on its website [24]. But the short mathematical description of the supported algorithms are given in the following parts.

2.7 Jarnac/JDesigner (2/11)

Jarnac/JDesigner [25, 26] is a script language for describing and manipulating the cellular system models. Here, Jarnac refers to a complicated script language and JDesigner stands for the design tool which can control the graphical interface of Jarnac in simulations of biochemical networks. By this way, Jarnac/JDesigner allows the user to describe metabolic, signal transduction and gene networks, or any physical system which can be represented in terms of a network and associated flows.

Additionally, Jarnac is written in Java and supports import and export of models from the SBML file format providing exchangeability of mathematical models of biological processes.

Furthermore, similar to previous tools, Jarnac is an open source software which is free for both academic and commercial purposes. Moreover it has a textual interface which requires the description of the model in the textual form. JDesigner, on the other hand, adds a diagrammatic interface on Jarnac.

Hence, the functionalities of Jarnac/JDesigner can be stated as follows [27, 28]:

- **Modeling Environment**: Jarnac is a modeling environment which is a better option for large-scale models. However, it is an advanced scripting language.
- **Windows Based**: Jarnac/JDesigner is only available on the Windows platform.
- **Built-in Computational Support**: The dynamic simulation (LSODA or CVODE integrator), steady-state analyses (NLEQ solver), simple stability analyses (eigenvalues analyses), matrix arithmetic (using the `IMSL` library), metabolic control analyses (all steady-state control coefficients and elasticizes), metabolic structural analyses (null space and conservation relation analyses with others to follow) and the stochastic simulation (Gillespie method) can be performed by this tool.

Lastly, Jarnac/JDesigner has a detailed manual and tutorial which can be found on its website [27, 28] and similar to the previous tools, the supported algorithms via Jarnac/JDesigner are explained mathematically in the following parts.

2.8 Systems Biology Toolbox for MATLAB (2.1)

Systems Biology Toolbox [29] is a toolbox for MATLAB to model and simulate biological and biochemical systems. This toolbox presents a user extensible context where new methods and applications as well as the simulation can be built. Moreover, it can provide the network identification, sensitivity and bifurcation analyses. On the other side, it has no design tool. Thus, JDesigner can be used for this manner [26].

Systems Biology Toolbox supports import and export of models from the SBML file format providing the exchangeability of mathematical models for biological processes with the SBML Toolbox. The user can also import and export the CSV data files and export to Maple or C/C++ files by writing the required codes.

Furthermore, it is designed to be in the MATLAB environment. It has a textual interface and allows two different representations of equations such that the first one is based on biochemical reaction equations, suitable for biochemists and biologists, and the second one depends on differential equations. In this toolbox, both representations are interchangeable.

Finally, its major functionalities can be listed as the following way [30].

- **Flexible**: By means of a textual interface, the users are not limited by the tool, rather, they can extend the functionality of the tool according to their needs.
- **Multi-platform**: Systems Biology Toolbox is a multi-platform tool since it is designed in MATLAB. Therefore, it is supported by other platforms such as Windows, Unix/Linux, and Mac OS X. However, the stochastic simulation is only available on the Microsoft Windows platform.
- **Different Functions**: Systems Biology Toolbox offers the three types of functions, namely, the auxiliary functions, information and editing functions and the analyses' functions. The last functions include the functions for deterministic and stochastic simulations and a more in silico experiment-oriented type of simulation. Moreover, the user can choose one of the seven deterministic solvers (Runge–Kutta, Adams, NDFs (BDFs), Rosenbrock, Trapezoidal Method, TR-BDF2, BDFs) and one of the three common stochastic algorithms (Direct method for exact simulation and Binomial τ-leap (also called as Binomial time-step) and Poisson τ-leap (also called as Poisson time-step) algorithms for approximate simulation).
- **Inference of Model Parameters**: Systems Biology Toolbox applies several optimization methods in the deterministic simulations and the estimation of the associated model parameters. These methods can be listed as a nonlinear solver based on the Newton iterations, local and global optimization functions based on the Nelder–Mead downhill simplex and the simulated annealing approaches.
- **Bifurcation Analysis**: The toolbox also provides a function for performing the bifurcation analysis with the help of a third party software XPP (XPPAUT) which is for free. The bifurcation happens when there is a parameter that causes a qualitative change in the dynamics of the system.

In the end, Systems Biology Toolbox has a manual and detailed tutorial which can be found on its website [30].

2.9 Virtual Cell (5.2 Beta)

Virtual Cell [31] indicates an internet simulation environment for the modeling and simulation of the cell biology. It has been specifically designed to be a tool for a wide range of scientists from experimental cell biologists to theoretical biophysicists [32].

Virtual Cell uses a web-based Java interface to specify the compartmental topology and geometry, molecular characteristics, and relevant interaction parameters. On the other side, the new users are supposed to be registered when they first run the Virtual

Cell Software. Additionally, Virtual Cell is an open source software which is free for both academic and commercial applications. It has a diagrammatic interface enabling the user to draw diagrams in place of writing chemical equations manually. Accordingly, the rate expressions and parameters of the selected systems can be entered by dialog boxes.

Hence, the main functionalities of Virtual Cell can be itemized as below [32]:

- **Collaborative Work**: Since Virtual Cell is an internet simulation server, it allows multiple researches at distant locations to collaborate in the development and analysis of a model. It also enables them to reuse, update, publish and privately share models amongst collaborating groups.
- **Multi-platform**: Virtual Cell is supported by all JAVA enabled platforms.
- **Different Simulation Methods**: Virtual Cell provides a collection of deterministic (Forward Euler (First Order, Fixed Time Step), Runge–Kutta (Second Order, Fixed Time Step), Runge–Kutta (Fourth Order, Fixed Time Step), Adams–Moulton (Fifth Order, Fixed Time Step), Runge–Kutta–Fehlberg (Fifth Order, Variable Time Step), IDA (Variable Order, Variable Time Step, ODE/ DAE), CVODE (Variable Order, Variable Time Step), Combined stiff solver CVODE/ IDA) and stochastic methods (Gibson (Next Reaction Stochastic method)) as well as hybrid methods (Hybrid (Gibson and Milstein method), Hybrid (Adaptive Gibson and Milstein method)) for solving the dynamics of a model. It also uses the CVODE solver for the optimization in the calculation of the parameter estimations.
- **Inference of Model Parameters**: Virtual Cell implements the particle swarm optimization algorithm with a simulated annealing based on the local search engine (LPEPSO-SA) in the calculations where necessary in the estimation of network parameters deterministically.

Lastly, Virtual Cell offers a detailed manual and tutorial which can be detected on its website [32] and the brief mathematical explanation about the supported algorithms is given in the following parts.

3 Deterministic and Inference Algorithms

In this part, we briefly present all the supported deterministic and optimization methods in the simulation tools above to explain their limitations and advantages.

3.1 Deterministic Algorithms

In general, the deterministic description is mainly described by the systems of the ordinary differential equations (ODEs) in biological systems in such a way that each differential equation denotes the rate of change in the concentration of the species and

consists of the concentration of these species. Therefore, they design numerical or exact methods to solve ODEs and understand the dynamics of the biological systems. Hereby, in this part, in order to define the capacity of each tool better, we represent short mathematical contents of the ODE methods supported by our simulation tools.

3.1.1 Euler Methods

The ODE systems can be presented as a vector of ODE specifying the relationship between a dependent variable $y = y(t)$ and an independent variable x with an initial condition which is mainly the initial amount for species (IVP) as in Eq. 1:

$$\frac{dy}{dt} = f(y,t) \ , \ y(t=0) = y(t_0) = y_0. \tag{1}$$

In this expression, t denotes the time and y_0 indicates the initial amount of y. Then by using the slope or the derivative of y, at the given time step $t = t_i$, by taking a step h towards the future realization of a given system, the Euler method can be formulated as

$$y_{i-1} = y_i + hf(y_i, t_i) \tag{2}$$

where $t = t_i$ $(i = 0, 1, \ldots)$ and h stands for the step-size. The forward Euler method comes from the truncated Taylor series expansion. Thereby, if y is extended in the neighborhood of $t = t_{i+1} = t_i + h$, the following equation can be obtained:

$$y(t_i + h) \equiv y_{i+1} = y(t_0) + h\frac{dy}{dt}\,|_{t_i} + O(h^2) = y_i + hf(y_i, t_i) + O(h^2), \tag{3}$$

in which $O(h^2)$ is the local truncation error (LTE).

The local truncation error (LTE) is reduced with each time-step because of the Taylor series expansion. In the forward Euler method, the LTE is denoted as $O(h^2)$. Therefore, the method is considered as a first order method. On the other hand, the global error g_i is the absolute value of the difference between the true solution and the computed solution, i.e., $g_i = |y^e(t_i) - y_i|$. Since mostly the exact solution cannot be known, calculation of the global error is not possible. However, it can be assumed that the global error at the ith time step is i times the LTE. So as i is proportional to $1/h$, the global error, g_i, is taken as proportional to LTE/h which means that for a kth order method, the global error becomes h^k [33, 34].

Furthermore, the forward Euler method is an explicit method, which is easy to implement and can be comparable with the implicit method. However, the drawback arises from the limitations on the time step size to ensure the numerical stability.

Thereby, the implicit method, also known as the backward Euler method, is formulated in order to obtain a numerically stable result as follows:

$$y_{i-1} = y_i + hf(y_{i-1}, t_{i+1}), \tag{4}$$

in which $f(y_i, t_i)$ is the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$ and h shows the step-size as used previously. Accordingly,

$$y_i \equiv y(t_{i+1} - h) = y(t_{i+1}) - h\frac{dy}{dt} \mid_{t_{i+1}} + O(h^2), \tag{5}$$

where $O(h^2)$ is the local truncation error (LTE). Among the listed simulation tools, Cellware, E-CELL, GENESIS and Virtual Cell support this method in the deterministic simulations of the biochemical systems.

3.1.2 Runge–Kutta Methods

In order to increase the accuracy of the Euler method, the Runge–Kutta methods are proposed by using the information on more than one point. The Runge–Kutta methods are popular and supported by all of the listed simulation tools in this study.

The most common orders of these approaches can be presented as below [33]:

i. *The Second-Order Runge–Kutta Method (Midpoint or Heun's Method):*
This method improves the Euler method by adding a midpoint in the step which increases the accuracy by one order. In other words, a better slope can be obtained by having the average of the two slopes coming from the Euler method as shown below via k_1 and k_2:

$$k_1 = hf(y_i, t_i), \tag{6}$$
$$k_2 = hf(y_i + k_1, t_i + h), \tag{7}$$

where $f(y_i, t_i)$ shows the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$. Furthermore, h denotes the step-size as given before. Hereby, the function y can be approximated by

$$y_{i+1} = y_i + (k_1 + k_2)/2 \tag{8}$$

by using k_1 and k_2. Finally, the second order Runge–Kutta method is an explicit method and is only conditionally stable in such a way that under that condition, the local truncation error LTE is found as $O(h^3)$.

ii. *The Fourth-Order Runge–Kutta Method:*
One of the most widely used deterministic methods is the fourth order Runge–Kutta approach which applies a weighted average of four slopes denoted by k_1, k_2, k_3 and k_4 as written below or the derivative of y:

$$k_1 = hf(y_i, t_i) \tag{9}$$
$$k_2 = hf(y_i + k_1/2, t_i + h/2) \tag{10}$$
$$k_3 = hf(y_i + k_2/2, t_i + h/2) \tag{11}$$
$$k_4 = h(y_i + k_3, t_i + h) \tag{12}$$

where $f(y_i, t_i)$ is the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$ while h presents the step-size as previously described. Hence, the method can be summarized by the following expression.

$$y_{i+1} = y_i + (k_1 + 2k_2 + 2k_3 + k_4)/6. \tag{13}$$

Finally, similar to the second-order Runge–Kutta method, the fourth-order Runge–Kutta approach is an explicit method and is conditionally stable while LTE is $O(h^5)$. More mathematical details about these methods can be found in [34].

3.1.3 Adams Methods

The Adams methods are based on approximating the integrand with a polynomial within the interval (t_i, t_{i+1}) as shown in the formula below. There are both explicit and implicit techniques within this method. The explicit technique is called as the Adams–Bashforth (AB) approach while the implicit one is called as the Adams–Moulton (AM) technique [33]:

$$y_{i+1} = y_i + \int_{t_i}^{t_{i+1}} \frac{dy}{dt} dt = y_i + \int_{t_i}^{t_{i+1}} f(y, t) dt \tag{14}$$

in which $f(y_i, t_i)$ describes the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$.

The first order AB method is equivalent to the forward Euler method while the AM method is equivalent to the backward Euler method. The mathematical details of both approaches are presented below. On the other hand, among the given simulation tools, Cellware, Jarnac/JDesigner, Systems Biology Toolbox and Virtual Cell perform these approaches in their deterministic calculations.

i. *Adams–Bashforth Method:*
The AB method is an explicit technique which is conditionally stable. The popular second order version is AB2 and is obtained by performing a linear interpolation. Accordingly, it can be formulated as follows:

$$y_{i+1} = y_i + \frac{h}{2}\left(3f(y_i, t_i) - f(y_{i-1}, t_{i-1})\right), \tag{15}$$

where $f(y_i, t_i)$ is the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$ as well as $t = t_{i-1}$. Finally, h stands for the step-size as used beforehand. Here, different from the Euler method, the solution from the $(i-1)$th and the ith steps are required to find the solution at the $(i+1)$th step.

ii. *Adams–Moulton Method (Trapezoidal Rule):*
The AM method is an implicit and stable method. The popular second order version is AM2 and is given by the following equation.

$$y_{i+1} = y_i + \frac{h}{2}\left(f(y_{i+1}, t_{i+1}) + f(y_i, t_i)\right), \tag{16}$$

in which $f(y_i, t_i)$ shows the derivative of the function at $y = y_i$ and $t = t_i$ $(i = 0, 1, \ldots)$ as well as $t = t_{i-1}$ when h implies the step-size. In general, this method is more costly in the computation as compared to AB. However, this is the main trade-off between the stability and the computational cost since both AM2 and AB2 have the second order accuracy.

3.1.4 Predictor - Corrector Methods

These methods combine the explicit and implicit deterministic techniques in order to obtain a method with better convergence characteristics. The most well-known combination is the combination of the forward Euler and the Adams–Moulton method. In this method, the forward Euler equation is used as a predictor equation to get a predictor for the $(i+1)$th step, y^p_{i+1}, and the AM2 equation is implemented as a corrector equation in order to obtain the final solution for the $(i+1)$th step, y_{i+1} [33]. This technique is called as the Euler-Trapezoidal method which can be computed as follows [35]:

$$y_{i+1} = y_i + hf(y_i, t_i), \tag{17}$$

$$y_{i+1} = y_i + \frac{h}{2}\left[f(y^p_{i+1}, t_{i+1}) + f(y_i, t_i)\right], \tag{18}$$

where $f(y_i, t_i)$ describes the derivative of the function at $y = y_i$, $t = t_i$ $(i = 0, 1, \ldots)$, $t = t_{i-1}$ and h indicates the step-size. Among the listed simulators, these methods are supported by Cellware, COPASI and Virtual Cell.

3.1.5 CVODE

CVODE [36] is a package for solving the initial value problems given in the explicit form $y' = f(t, y)$. The package is written in C and can handle both stiff and non-stiff systems. Specifically, when numerical methods for solving the differential equations are numerically unstable and when the step-size is not extremely small, then it is considered as a stiff case. Otherwise, the system is accepted as a non-stiff manner. For the nonstiff cases, the Adams–Moulton method is performed with the possible order varying between 1 and 12 while for the stiff cases, the Backward Differentiation Formulas (BDFs) with the possible order varying between 1 and 5 are applied. In both methods, the system is solved at each integration step. To do this, CVODE presents a functional iteration which is a good selection for only non-stiff systems and the diverse method of the Newton iteration.

Among tools, Jarnac/JDesigner and Virtual Cell implement this method in their computations.

3.1.6 LSODA

LSODE (Livermore Solver for Ordinary Differential Equations) is a package for solving initial value problems given in an explicit form $y' = f(t, y)$. The package is written in FORTRAN. For the non-stiff cases, the Adams–Moulton method is implemented with the possible order varying between 1 and 12 while for the stiff cases, the Backward Differentiation Formulas (BDFs) with the possible order varying between 1 and 6 are computed. Moreover, with LSODA, the users do not have to determine whether the problem is stiff or not. Hence, the solver can automatically choose the appropriate method with starting from non-stiff methods [37].

Similarly, this package is supported by the Jarnac/JDesigner and Virtual Cell simulation tools.

3.2 Inference Algorithms

In the concept of the biochemical modeling, the inference of the model parameters supported by the simulation tools is performed deterministically via optimization methods. The most widely used optimization approaches for the estimation in these tools can be listed as follows.

3.2.1 Evolutionary Algorithms (EA)

The evolutionary algorithm (EA) is a heuristic optimization algorithm hold on the reproduction and selection mechanism to find an optimal solution for a biochemical system having specific constraints. In this algorithm, each individual in the system is a candidate solution to the problem and is represented by a genome in which each gene subtends to a parameter. When generating the algorithm, each individual reproduces two different individuals. One of them is the same with the parent and the other is exposed to some mutations. Later, each individual is ranked by the number of wins (i.e., the number of individuals representing worse solutions than it). In the end, half of the individuals with the highest ranks is remained so that the population again has the original number of the individuals [7].

This algorithm has three subcategories, namely, gene expression programming, genetic algorithms and genetic programming. Among them, the genetic algorithm is the most common approach. In this algorithm, at each generation, individuals again reproduce two offsprings. However, these offsprings are produced by combining the genomes of their parents [7].

In evolutionary algorithms, the stochastic ranking can also be used in place of the usual ranking and here, it is specifically named as the evolutionary strategies with stochastic ranking (SRES) and the genetic algorithm with stochastic ranking (SR).

Lastly, thanks to the availability of working parallel computers, the tools can support parallel evolutionary algorithms [22] which helps the user to apply genetic algorithms to large populations. As advantages of parallel algorithms, robustness, easy customization for a new problem, and multiple-solution capabilities can be listed.

Among the simulation tools, COPASI and GENESIS support these algorithms.

3.2.2 Particle SWARM Algorithm (PSA)

The particle SWARM algorithm is a population based stochastic optimization method developed by Kennedy and Eberhart in 1995, inspired by social behavior of bird flockings or fish schoolings [38].

In this algorithm, the particles having a position X_i and a velocity V_i in the parameter space, remember their best achieved objective values O and positions M_i. Given this information and the position of their best neighbor, which is a random subset of particles of the swarm, a new velocity is calculated. Then, the position is updated [7].

The particle SWARM Algorithm is similar to the Evolutionary Algorithms such as Genetic Algorithms. However, PSA has no evolution operators such as crossover and mutation. Instead, the candidate solutions lie through the problem space by following the current optimum particles.

Similarly, among tools, Cellware, COPASI and Virtual Cell offer this method for the optimization.

3.2.3 Simulated Annealing (SA)

The simulated annealing is a heuristic stochastic optimization method which can converge to the global optimum of the objective function if the number of iterations goes to infinity. This method is robust whereas, it is slow compared to the other global optimization algorithms.

In SA, the objective function is kept constant. During each iteration, the parameters are changes randomly within the parameter space and the new objective function value is computed. If the computed value is less than the previous value, then the new state is accepted. Otherwise, the new state is accepted with a probability coming from the Boltzmann distribution. After a fixed number of iterations, the stopping criterion of the algorithm is checked. If it is not maintained, then the system's temperature is reduced and the algorithm continues until the criterion is satisfied [7].

Lastly, since this method is a robust, but, a slow algorithm, parallel simulated algorithms are also applicable in some tools such as COPASI, GENESIS, Systems Biology Toolbox and Virtual Cell.

3.2.4 Hooke and Jeeves (Pattern Search - PS) Algorithm

This algorithm is a direct search optimization method which minimizes a nonlinear function without using the derivative information. Hooke and Jeeves algorithm is heuristic in the sense that it suggests a descent direction using the values of the function calculated in a number of previous iterations. This algorithm is also considered as a derivative-free method since it can work with the functions that are not continuous or differentiable [7]. Among tools, COPASI performs this method.

3.2.5 Levenberg–Marquardt Algorithm

This method is a combination of the steepest descent and the Newton methods. Therefore, it is a gradient descent method. The Newton optimization method follows descent directions computed from the first and the second partial derivatives and minimizes a nonlinear function. On the other hand, the steepest descent method only applies the first derivative of the function. However, it can guarantee to the convergence unlike the Newton method [7]. Similarly, COPASI uses this approach in the optimization.

3.2.6 Nelder–Mead Algorithm (Downhill Simplex Method)

The Nelder–Mead method is a nonlinear optimization algorithm proposed by Nelder and Mead (1965) which minimize a nonlinear function of several variables without needing the information of derivatives.

In this method, when there are N variables, the simplex is a polytope of $(N + 1)$ vertices. To illustrate, when there are two variables, a simplex is a triangle and the method is based on a pattern search comparing the values of the objective function at each vertex. Later, the worst one is replaced with a point reflected through the centroid of the other N points. If this point is better than the best current point, then it is tried to stretch exponentially out along this line. If this new point is not much better than the previous value, then it is stepped across a valley so that we can shrink the simplex towards the best point [7]. Finally, it is supported by COPASI and Systems Biology Toolbox among the listed simulation tools.

3.2.7 Random Search (RS) Algorithm

The random search algorithm is a numerical optimization method which can minimize an objective function without needing the gradient. In this method, a series of combinations of random values of the parameters is generated. Later, the generated random values not satisfying the constraints are removed. As the number of iterations becomes very large, then the global optimum of the objective function can be found [7]. Finally, this method is used in the COPASI tool during the optimization.

4 Stochastic Algorithms

The dynamical behavior of biochemical systems can be also analyzed via stochastic methods. Unlike the deterministic methods, these approaches consider a random error term coming from the Brownian motion.

Mathematically, the main differences between the deterministic and stochastic approaches are the role of reaction constants and the description of the amount of species. Moreover, in the deterministic approach, the amounts of species are the concentrations while it is the number of molecules in the stochastic approach. Furthermore, in the stochastic approach, the dynamics of the system are governed by the chemical master equation (CME) [39] which is dependent on the reaction probabilities.

On the other hand, when all reactions are zero and the first-order mass action kinetics, the deterministic solution of the system correctly describes the expected value of the stochastic kinetic model. Hereby, in this part, the main exact and approximate stochastic simulation algorithms are introduced and the supported tools are listed in the associated parts [40].

4.1 Exact Stochastic Algorithms

4.1.1 Gillespie Algorithm (Direct Method)

The Gillespie algorithm [41] is one of the exact algorithms presented for stochastic simulations of biochemical systems. In the stochastic approach, since the reaction hazards depend merely on the current state of the system, the time evaluation of the state in the reaction systems becomes as a continuous time Markov process with a discrete state space. In stochastic modelling, the hazard is also called as the rate law of reaction and describes the product of the number of distinct molecular reactant combinations available in the state x for reaction i ($i = 1, \ldots, v$) with stochastic rate constant c_i. Hence, in a given reaction system with v reactions, the hazard for the ith reaction can be denoted as $h_i(x, c_i)$. Accordingly, the total hazard $h_0(x, c)$ for the system after the occurrence of a reaction is computed as

$$h_0(x, c) = \sum_{i=1}^{v} h_i(x, c_i). \tag{19}$$

Therefore, the time for the next reaction can have a distribution coming from the exponential with rate $h_0(x, c)$. Here, x and c denote the state, i.e. the number of molecules, and the reaction rate constant, respectively. Moreover, the underlying picked probabilities are proportional to $h_i(x, c_i)$ and independent on the time of the next event. In other words, the reaction type is found by the probability $h_i(x, c_i)/h_0(x, c)$. Then, by using the time to the next event and the event time, the state of the system can be updated and the simulation continues [40].

Hence, the step of this algorithm can be stated as below:

(i) Initialize the system at time $t = 0$ with stochastic reaction rate constants $c_1, c_2 \ldots, c_v$ and the numbers of molecules for each species $x_1, x_2 \ldots, x_u$.
(ii) For each $i = 1, \ldots, v$ among a system with v reactions, update the hazard function $h_i(x, c_i)$ based on the current state, x.
(iii) Calculate the total hazard via $h_0 = \sum_{i=1}^{v} h_i(x, c_i)$.
(iv) Simulate the time of the next event by $t' = -\ln[U(0, 1)]/h_0$, where $U(0, 1)$ denotes the standard uniform random number and puts this in the update of the current time via $t \equiv t + t'$.
(v) Simulate the reaction index, j as a discrete random quantity with probabilities $h_i(x, c_i)/h_0(x, c)$ where $i = 1, \ldots, v$.
(vi) Update v according to the reaction j by putting the current state x by $x \equiv x + s^{(j)}$, where $s^{(j)}$ denotes the jth column of the stoichiometry matrix s.
(vii) If the total simulation time T is exceeded after the update of the time, then stop. Otherwise, go back to Step (ii).

Among tools, Cellware, Dizzy, Dynetica, GENESIS, Jarnac/JDesigner and Systems Biology Toolbox use this method for the stochastic simulations.

4.1.2 First Reaction Method

This algorithm can be considered as the extension of the direct method in the sense that it depends on the generation of the next time step via the shortest time hazard from the exponential distribution, in place of the total hazard from the same density [40]. Thereby, the step of the algorithm can be summarized as follows.

(i) Initialize the starting point of the simulation with the time $t = 0$, the reaction rate constants $c = (c_1, c_2, \ldots, c_v)$ for totally v reactions and the number of molecules for each species $x = (x_1, x_2, \ldots, x_u)$ for totally u species in a system.
(ii) Calculate the reaction hazards $h_i(x, c_i)$, where $i = 1, \ldots, v$.
(iii) Simulate a putative time to the next reaction from the exponential density with a rate h_i, i.e., $t_i \sim \exp(h_i(x, c_i))$
(iv) By denoting j as the index of the smallest t_i, put the time t as $t \equiv t + t_j$. Here t_j denotes the time step.
(v) Update the state according to the reaction with an index j and set the state x to $x \equiv x + s^{(j)}$.
(vi) If the updated t is less than the total simulation time T, then return to Step (ii). Otherwise, stop the simulation.

Among tools, GENESIS uses this method for stochastic simulations.

4.1.3 Gibson–Bruck Algorithm (Next Reaction Method)

Gibson and Bruck (2000) propose a new simulation approach in order to reduce the computational cost of the Gillespie algorithms since Gillespie becomes very demanding in terms of the calculational time under the simulation of large systems. Accordingly, this method can be considered as a modification of the first reaction method which is computationally more efficient [40]. Hereby, the step of this algorithm can be summarized subsequently as below:

(i) Initialize the system at time t=0 with the rate constants $c_1, c_2, \ldots, c_v$ and the initial numbers of the molecules for each species $x_1, x_2, \ldots, x_u$ for a system with u species and v reactions. Then, calculate all of the initial hazards $h_i(x, c_i)$ where $i = 1, \ldots, v$. These hazards are used to simulate the putative time for the first reaction times t_i from the exponential density with a rate $h_i(x, c_i)$.
(ii) Considering that j is the index of the smallest t_i, set the time t to $t = t_j$.
(iii) Update the state x according to the reaction with the index j.
(iv) Update the hazard under the state x and the rate constant c, i.e., $h_j(x, c_j)$, according to the new state x. Then, simulate a new putative time t_j via $t_j = t + \exp(h_j(x, c_j))$ from an exponential increment with rate $h_j(x, c_j)$.
(v) For each reaction $i \neq j$ whose hazard is changed by the reaction j, update the hazard and the time via $h_i' = h_i(x, c_i)$ and $t_i = t + (h_i / h_i')(t_i - t)$, respectively.

(vi) If the updated time is less than the total simulation time, i.e., $t < T$, return to Step (ii).

Different from the previous algorithms, the Gibson–Bruck method works with absolute times (i.e. the time to the next event), rather than relative times (i.e. times from now until the next event). By this way, it saves times during the simulation of the new time steps for all of the reactions which are not affected by the putative reaction. Moreover thanks to this idea, the times which are not affected by the most recent reaction is reused by appropriately rescaling the old variable. Furthermore, it is assumed that the algorithm knows which hazards are affected by each reaction that can be done by creating a dependency graph for the system. Using this graph, if a reaction of type j occurs, the set of all children of the node j in the graph gives the set of hazards that needs to be updated. An interesting alternative to this graph can be a direct implementation of the Petri net representation [40].

All in all, this algorithm is more efficient than the Gillespie's direct method since only one new random number needs to be simulated for each reaction, compared to Gillespie in which two random numbers are required [8, 40].

Among tools, Cellware, COPASI, Dizzy, GENESIS and Virtual Cell use this method for the stochastic simulations of the biochemical systems.

4.2 Approximate Stochastic Simulation Algorithms

If the user disregard the exactness of the simulation, then the computation cost for the simulation of the system can be reduced significantly. In order to perform such approximated generation of the systems,the following methods can be applied among many alternatives in the literature of the approximate stochastic simulation methods.

4.2.1 Time Discretization Method (Poisson Time-Step Method)

In this method, the time axis is divided into the small discrete parts, and the underlying kinetics are approximated so that the advancement of the state from the start of one part to another can go on. In other words, it is assumed that the time intervals are sufficiently small that the reaction hazards can be constant over the interval. This condition is also known as the leap condition. Thus, in this approach, the number of reactions occurring in a short time interval is assumed to come from the Poisson distribution. Later, numbers of reaction events can be simulated and the system can be updated [40].

As a result, for a fixed (small) time step Δt, we can represent an approximate simulation algorithm as follows.

(i) Initialize the system at time $t = 0$ with the rate constants $c_1, c_2, \ldots, c_v$ and the initial numbers of the molecules for each species $x_1, x_2, \ldots, x_u$ as well

as the stoichiometry matrix S whose entry can be denoted by s_r for the rth ($r = 1, \ldots, v$) reaction.

(ii) Calculate the hazard of each reaction $h_i(x, c_i)$ for $i = 1, \ldots, v$ and simulate the u-dimensional reaction vector r, with the ith entry from a Poisson density via $\text{Poi}(h_i(x, c_i)\Delta t)$.

(iii) Update the state according to $x \equiv x + s_r$.

(iv) Update the current time t via $t \equiv t + \Delta t$ for the change of time Δt.

(v) If the current time is less than the predefined total time T, i.e., $t < T$, go back to Step (ii).

In this approach, the crucial point is to choose a convenient time step Δt so that the method is fast, but, also accurate enough. Because the smaller Δt, the more accurate and the larger Δt, the faster simulation can be done. Moreover, suitability of a specific Δt can change in each part of the simulation. Therefore, instead of considering a constant Δt, τ-leap method is proposed which considers a variable time-step Δt [40, 42]. Among tools, E-CELL and Systems Biology Toolbox use this method for the approximate stochastic simulations.

4.2.2 Gillespie's Time-Step (τ-Leap) Method

The time step, also known as τ-leap method [42], is a version of the Poisson time-step method which enables going forward by a variable amount τ, in which τ balancing the trade-off between the accuracy and the speed. In this sense, τ is selected as large as possible and satisfies some constraints for accuracy.

More specifically, the notion of this method is to choose τ in such a way that the proportional change in all of the hazard, $h_i(x, c_i)$, is small. To achieve this, the easiest method can be a post-leap check. In other words, after a leap τ, the expression $|h_i(x', c_i) - h_i(x, c_i)|$ can be checked if it is sufficiently small or not. If it is large, then the search for a smaller value of τ should continue [40].

On the other hand, a pre-leap method might be better if the expected new state can be written as $E(x') = x + E(r)A$ where the ith element of $E(r)$ is $h_i(x, c_i)\tau$. Here, $E(.)$ describes the expected value of the underlying random variable. Then, to check whether the change in the hazard at this expected new state is sufficiently small, the following inequality can be used:

$$|h_i(x', c_i) - h_i(x, c_i)| \leq \varepsilon h_0(x, c). \tag{20}$$

In this expression, ε presents a small interval, h_i and h_0 stand for the hazard of the ith reaction and the total hazard of the system, respectively. Here, Gillespie suggests an approximate method for computing the largest τ satisfying this property, i.e. the leap condition. Accordingly, if the resulting τ is as small as the expected time leap associated with an exact single reaction update, then it is preferable to simulate the system. Since the time of the first event is distributed as exponential with a rate

$h_0(x, c)$, i.e., $\exp(h_0(x, c))$, which has the expectation $1/h_0(x, c)$, we can always prefer an exact update if the suggested τ is less than say $4/h_0(x, c)$ for the total hazard h_0 under the state x and the reaction rate constants c [40].

Finally, among the major simulation tools, Cellware, Dizzy and Systems Biology Toolbox can apply this approach in the approximate stochastic simulations.

5 Hybrid Algorithms

Depending on the nature of a biochemical system studied, it is sometimes convenient to consider stochastic algorithms, especially, when the concentrations are low or the randomness has an important role. Otherwise, the traditional methods, namely, deterministic algorithms can be applied. However, it is also convenient to combine these two algorithms under a hybrid approach.

Thereby, in order to apply a hybrid algorithm, the entire biochemical system is divided into the two groups, called as deterministic and stochastic subnets. As it is mentioned beforehand, the deterministic subnet covers the reactions with species having high particle numbers. Accordingly, the hybrid method performs the deterministic numerical integration of ordinary differential equations in order to combine the deterministic method with a stochastic simulation algorithm. For our listed simulation tools, Cellware, COPASI, E-CELL and Virtual Cell support these algorithms.

6 Conclusion

In this study, we have presented the most widely used simulation tools designed for the biological networks. These tools can offer not only algorithms for simulation, both also algorithms for the inference of model parameters based on deterministic modelling. Furthermore, some of them can be able to make a variety of analyses for metabolite structure and bifurcation of the systems. Regarding the variety of the computational choices offered to the researchers, it can be seen that COPASI and Virtual Cell are particularly more comprehensive and Cellware is relatively comprehensive among alternatives.

As the extension of this study, we consider to define some criteria in order to compare all these approaches formally and show the performance of the best choice in a real-life example. Furthermore, all these tools support deterministic approaches. Hereby, we think to investigate their capacities under stochastic and impulsive differential equations as well. Moreover, we think to check whether these tools can be also supported by recent approximation methods such as the secant and Kurchatov methods [43] to solve the nonlinear expressions with high computational efficiency. Finally, we investigate the capacity of all tools in terms of modelling such as the modelling based on the gene-environment networks [44] or complex regression models [45, 46] which can deal with both stochastic and deterministic approaches.

Acknowledgements Vilda Purutçuoğlu would like to thank the Middle East Technical University, Academic Development Programme's Project (AGEP Project No: BAP-08-11-2014-007) for its support.

References

1. Dhar, P., et al.: Cellware-a multi-algorithmic software for computational systems biology. Bioinformatics **20**(8), 1319–1321 (2004)
2. Hucka, M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4), 524–531 (2003)
3. Systems Biology Group.: Bioinformatics Institute, Singapore, Cellware Manual. http://www.bii.a-star.edu.sg/docs/sbg/cellware/Manual_3_0_1.pdf
4. http://www.bii.a-star.edu.sg/achievements/applications/cellware
5. Hoops, S., et al.: COPASI-a complex pathway simulator. Bioinformatics **22**(24), 3067–3074 (2006)
6. Mendes, P.: GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. Comput. Appl. Biosci. CABIOS **9**(5), 563–571 (1993)
7. COPASI Development Team.: COPASI Documentation Version 4.6 (Build 32) (2010). http://www.copasi.org/tiki-index.php?page=User_Manual
8. Gibson, M.A., Bruck, J.: Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. **104**(9), 1876–1889 (2000)
9. http://www.copasi.org/
10. Ramsey, S., Orrell, D., Bolouri, H.: Dizzy: stochastic simulation of large-scale genetic regulatory networks. J. Bioinf. Comput. Biol. **3**(2), 415–436 (2005)
11. Ramsey, S.: Dizzy User Manual ver 1.11.4 (2006). http://magnet.systemsbiology.net/software/Dizzy/docs/UserManual.html
12. http://magnet.systemsbiology.net/software/Dizzy
13. You, L., Hoonlor, A., Yin, J.: Modeling biological systems using Dynetica-a simulator of dynamic networks. Bioinformatics **19**(3), 435–436 (2003)
14. You, L.: Dynetica User Guide (version 0.1beta) (2002). http://people.duke.edu/~you/Dynetica/DyneticaUserGuide.htm
15. http://people.duke.edu/you/Dynetica_page.htm
16. Tomita, M., et al.: E-CELL: software environment for whole-cell simulation. Bioinformatics **15**(1), 72–84 (1999)
17. Takahashi, K., et al.: E-CELL 2: multi-platform E-Cell simulation system. Bioinformatics **19**(13), 1727–1729 (2003)
18. Takahashi, K., et al.: E-CELL system version 3: a software platform for integrative computational biology. Genome Inf. Ser. **19**(13), 294–295 (2003)
19. http://www.e-cell.org
20. Bower, J.M., Beeman, D., Hucka, M.: The GENESIS simulation system. The Handbook of Brain Theory and Neural Networks, vol. 19(13), pp. 475–478. MIT Press, Cambridge (2003)
21. https://www.ncbs.res.in/node/350
22. Collins, R.J., Jefferson, D.R.: Selection in massively parallel genetic algorithms. In: Proceedings of the 4th International Conference on Genetic Algorithms, pp. 249–256 (1991)
23. Azencott, R.: Simulated Annealing: Parallelization Techniques. Wiley, New York (1992)
24. http://www.genesis-sim.org/GENESIS
25. Sauro, H.M., Fell, D.A.: Jarnac: a system for interactive metabolic analysis. In: Animating the Cellular Map: Proceedings of the 9th International Meeting on BioThermoKinetics, 19(13), pp. 294-295 (2003)
26. Sauro, H.M., Fell, D.A.: JDesigner: A simple biochemical network designer (2001)
27. http://sbw.kgi.edu/software/Jarnac.htm

28. http://sbw.kgi.edu/sbwWiki/sbw/jdesigner
29. Schmidt, H., Jirstrand, M.: Systems biology toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics **22**(4), 514–515 (2006)
30. http://www.sbtoolbox.org/
31. Loew, L.M., Schaff, J.C.: The virtual cell: a software environment for computational cell biology. Trends Biotechnol. **19**(10), 401–406 (2001)
32. http://vcell.org
33. Zeltkevic, M.: Differential Equations Notes (1998). http://web.mit.edu/10.001/Web/Course_Notes/Differential_Equations_Notes
34. Defterli, Ö.: Modern Mathematical Methods in Modeling and Dynamics of Regulatory Systems of Gene-Environment Networks. Middle East Technical University, Institute of Applied Mathematics (2011)
35. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford University Press, Oxford (1987)
36. Cohen, S.D., Hindmarsh, A.C.: CVODE, a stiff/nonstiff ODE solver in C. Comput. Phys. **10**(2), 138–143 (1996)
37. http://people.sc.fsu.edu/jburkardt/f77_src/odepack/odepack.html
38. http://www.swarmintelligence.org
39. Van Kampen, N.G.: Stochastic Processes in Chemistry and Physics. Chaos, (1981)
40. Wilkinson, D.J.: Stochastic Modelling for Systems Biology. Chapman & Hall, CRC Mathematical and Computational Biology Series. Taylor & Francis, Boca Raton (2006)
41. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977)
42. Gillespie, D.T.: Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. **115**(4), 17161733 (2001)
43. Garu-Sánchez, M., Noguera, M., Gutiérrez, J.M.: Frozen iterative methods using divided differences: à la Schmidt-Schwetlick. J. Opitim. Theory.Appl. **160**, 931–948 (2014)
44. Defterli, Ö., Purutçuoğlu, V., Weber, G.W.: Advanced mathematical and statistical tools in the dynamic modelling and simulation of gene-environment networks. In: Zilberman, D., Pinto, A. (eds.) Modeling, Optimization, Dynamics and Bioeconomy, pp. 235–257. Springer, Berlin (2014)
45. Hastie, T., Tibshirani, R., Friedman, J.: The Element of Statistical Learning. Springer, New York (2001)
46. Weber, G.W., Batmaz, I., 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**(3), 371–400 (2012)